Computer Science and Application 计算机科学与应用, 2020, 10(9), 1549-1558
Published Online September 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.109163

Hans汉斯

基于正负类边界距离的多标签数据属性约简

纪思南

渤海大学数学系，辽宁 锦州
Email: 1972324430@qq.com

收稿日期：2020年8月18日；录用日期：2020年8月28日；发布日期：2020年9月4日

摘　要

本文首先将多标签分类问题分解为一系列单标签二分类子问题，每一个子问题对应一个标签。子问题的正类定义为具有该标签的样本，负类定义为不具有该标签的样本。在给定的属性子集下，计算出正负类样本之间的最小距离，即分类边界的最小距离。将子问题分类边界最小距离求和定义为依赖度函数，并将此依赖度函数作为属性子集重要度评价指标。然后建立了所提出依赖度函数关于属性子集的单调性，并通过最大化依赖度函数给出了属性约简的定义。最后，设计了一种基于正负类边界距离的属性约简算法，并在实际的多标签数据集上进行了实验，实验结果表明，所提约简算法能够建立合理的属性重要度排序，有效地去除冗余属性。

关键词

多标签分类，属性约简，边界距离，依赖度函数

Attribute Reduction for Multi-Label Data Based on Boundary Distance between Positive and Negative Classes

Sinan Ji

Department of Mathematics, Bohai University, Jinzhou Liaoning
Email: 1972324430@qq.com

Received: Aug. 18th, 2020; accepted: Aug. 28th, 2020; published: Sep. 4th, 2020

Abstract

In this paper, the multi-label classification problem is decomposed into a series of single-label binary classification sub-problems, each of which corresponds to one label. The positive class of the

文章引用: 纪思南. 基于正负类边界距离的多标签数据属性约简[J]. 计算机科学与应用, 2020, 10(9): 1549-1558.
DOI: 10.12677/csa.2020.109163

sub-problem is defined as the sample with the label, and the negative class is defined as the sample without the label. Under the given attribute subset, the minimum distance between positive and negative class samples is calculated, that is, the minimum distance of classification boundary. The sum of the boundary distances for all sub-problems is defined as dependency function, which is used as the evaluation of the importance of the attribute subset. The monotonicity of the proposed dependence function with respect to the attribute subset is established, and the definition of the attribute reduction is given by maximizing the dependence function. Finally, an attribute reduction algorithm based on positive and negative class boundary distance is designed, and the experiments are carried out on actual multi-label data sets. The experimental results show that the proposed algorithm can establish a reasonable ranking of attribute importance and effectively remove redundant attributes.

Keywords

Multi-Label Classification, Attribute Reduction, Boundary Distance, Dependence Function

1. 引言

近年来，在数据挖掘和机器学习领域中，多标签分类问题引起了广泛的关注。多标签分类和单标签分类的区别就是单标签分类中的样本只能与一个标签相关联，而多标签分类中的样本可能同时与多个类别标签相关联[1] [2] [3]。

在多标签数据中，有些属性可能是冗余的或与分类任务不相关的，冗余或不相关的属性会导致多标签分类器的分类性能较差。因此，在设计分类器前，需要减少冗余或不相关的属性[4]。属性约简就是在分类能力不变的前提下，删除其中冗余或不相关的属性，选择一个最优属性子集来提高分类器的分类性能，保留了数据集最有用的信息，使分类模型简洁，具有更好的泛化能力[5] [6] [7]。

属性约简是多标签分类重要的数据预处理过程，其目的是减少数据的维数，来提高数据处理速度[8]。2018 年，Lin 等人提出了一种新的模糊粗糙集模型，采用局部采样技术构造样本间的鲁棒距离，并建立了一种适用于多标签学习的属性约简算法[9]。2019 年，Wang 等人提出了一种基于信息论的多标签学习特征选择算法，他先定义了多标签信息熵和多标签互信息的新概念，然后，建立了一种缺失标签的多标签特征选择方法[10]。2020 年，Liu 等人通过设计类间判别和类内邻域识别来选择每个新到达的标签特征，提出了一种在流标签环境下的多标签特征选择方法[11]。

现有的属性约简算法，计算复杂度较高，而样本间距离运算简单，直观，所以我们在这篇论文中利用样本间距离，定义依赖度函数，使模型更加简洁，运算速度更快。我们首先将多标签分类问题转化为二分类问题，并介绍了正负类样本集，然后，求出正负类样本间距离，从而定义了依赖度函数。通过正负类样本间距离函数，证明了依赖度函数相对于属性子集是单调递增的。接着给出了属性约简定义，最后，设计了一种基于正负类边界距离的属性约简算法，来选择最优的多标签属性子集。为了验证该方法的性能，我们将其与现有的四种多标签特征选择算法进行了比较。实验结果表明，该方法能够有效地去除冗余属性。

论文的其余部分结构如下。在第一节中，回顾了多标签分类的基础知识、点到点的距离、集到集的距离和点到集的距离的定义。在第二节中，我们给出了属性约简的定义，设计了一种基于正负类边界距

离的属性约简算法。在第三节中，汇报了我们的算法在 9 个多标签数据集上实验的结果。在第四节中，对本文所得结论和实验结果进行总结，并提出了今后要研究的问题。

2. 基础知识

多标签数据表示为 $S=(U,A,L)$，其中样本集 $U=\{x_1,x_2,\cdots,x_n\}$ 是一个非空有限集，表示有 n 个样本实例。属性集 $A=\{a_1,a_2,\cdots,a_p\}$ 是一个非空有限集，每个属性 $a\in A$ 是从样本集 U 到实数集的函数，其中 $a(x)$ 表示为样本 $x\in U$ 在属性 a 上的值。标签集 $L=\{l_1,l_2,\cdots,l_q\}$ 是一个非空有限集，表示有 q 个标签。每个标签 $l\in L$ 被定义为从样本集 U 到集合 $\{0,1\}$ 的函数。假设样本 x 与标签 l 相关联，则 $l(x)=1$；否则，$l(x)=0$。

对于任意非空属性子集 $B\subseteq A$，假设样本点 x 和点 y 都是样本集 U 中的点，则点 x 和点 y 之间的距离定义为

$$d_B(x,y)=\sum_{a\in B}\left|a(x)-a(y)\right|.$$

对于每个非空属性子集 $B\subseteq A$，设集合 X 和集合 Y 是样本集 U 的子集，则集合 X 和集合 Y 之间的距离定义为

$$d_B(X,Y)=\min_{x\in X,y\in Y}d_B(x,y).$$

当集合 X 是单点集 $\{x\}$ 时，点 x 和点集 Y 之间的距离为

$$d_B(x,Y)=\min_{y\in Y}d_B(x,y).$$

任意两个非空集 X 和 Y 都有 $d_B(X,Y)\geq 0$。

3. 多标签数据属性约简

在给定标签 $l_i\in L$ 的情况下，关于 l_i 的一对样本集被定义为

$$E_i=\{x\in U:l_i(x)=1\},i=1,\cdots,q.$$

$$F_i=\{x\in U:l_i(x)=0\},i=1,\cdots,q.$$

样本集 E_i 是有该标签的样本集合，样本集 F_i 是没有该标签的样本集合，E_i 是正类样本集，F_i 是负类样本集。每个标签具有一对样本集 E_i 和 F_i。对于给定的属性子集 $B\subset A$，下面的函数 $d_B(E_i,F_i)$ 可以测量正类样本集 E_i 和负类样本集 F_i 之间的距离，函数定义为

$$d_B(E_i,F_i)=\min_{x\in E_i,y\in F_i}d_B(x,y).$$

这个距离函数 $d_B(E_i,F_i)$ 测量了正类样本集 E_i 和负类样本集 F_i 的边界距离，函数值越大，分类器的分类能力越好。也就是说，标签相同的样本距离越近越好，标签不同的样本距离越远越好。

定义 1：对于任意非空属性子集 $B\subseteq A$，我们将每个标签的正负类样本间距离求和定义为依赖度函数，依赖函数为

$$\gamma(B)=\sum_{i=1}^{q}d_B(E_i,F_i)$$

依赖度函数 $\gamma(B)$ 用于评估属性子集的重要性，依赖度函数值的大小取决于正负类样本间距离函数 $d_B(E_i,F_i)$ 值的大小。

引理 1：对于多标签数据 $S=(U,A,L)$，距离函数 $d_B(E_i,F_i)$ 和依赖度函数 $\gamma(B)$ 相对于属性集 B 是单调递增的。任意选择两个属性子集 B_1 和 B_2，有 $B_1\subset B_2\subset A$，则存在 $d_{B_1}(E_i,F_i)<d_{B_2}(E_i,F_i)$，$\gamma(B_1)\leq\gamma(B_2)$。

证明：由距离的定义，对于任意 $B_1 \subset B_2$，有

$$d_{B_1}(x,y)=\sum_{a\in B_1}\left|a(x)-a(y)\right|\le\sum_{a\in B_2}\left|a(x)-a(y)\right|=d_{B_2}(x,y)$$

根据集合到集合间距离的定义有

$$d_{B_1}(E_i,F_i)=\min_{x\in E_i,y\in F_i}d_{B_1}(x,y)\le d_{B_1}(x,y)$$

$$d_{B_2}(E_i,F_i)=\min_{x\in E_i,y\in F_i}d_{B_2}(x,y)\le d_{B_2}(x,y)$$

因此

$$d_{B_2}(E_i,F_i)-d_{B_1}(E_i-F_i)>0$$

$$d_{B_1}(E_i,F_i)<d_{B_2}(E_i,F_i)$$

另外，根据定义 1 有

$$\gamma(B_1)=\sum_{i=1}^{q}d_{B_1}(E_i,F_i)$$

$$\gamma(B_2)=\sum_{i=1}^{q}d_{B_2}(E_i,F_i)$$

得到

$$\sum_{i=1}^{q}d_{B_2}(E_i,F_i)-\sum_{i=1}^{q}d_{B_1}(E_i,F_i)\ge 0$$

因此

$$\gamma(B_1)\le\gamma(B_2)$$

引理得证。

从引理 1 看出，随着属性数目的增加，样本间距离函数值越大，依赖度函数值也越大。这意味着在属性子集中添加新属性从而提高了分类器的分类能力。现在，我们通过最大化依赖度函数给出属性约简的定义。

定义 2：对于多标签数据 $S=(U,A,L)$，如果属性子集 $B\subseteq A$ 满足以下条件，则称集合 B 是 A 的依赖约简：

1) $\gamma(B)=\gamma(A)$；

2) 对于任意的 $B'\subset B$，$\gamma(B')\ne\gamma(A)$。

我们可以看到约简的是 A 的一个极小子集，它保持了分类器的分类能力。然而，在现实应用中，上述定义过于严格，因此我们引入一个参数 ε，并给出基于正负类边界距离的属性约简定义。

定义 3：对于多标签数据 $S=(U,A,L)$，如果属性子集 $B\subseteq A$ 满足以下条件，则称集合 B 是 A 基于正负类边界距离的属性约简：

1) $\left|\gamma(B)-\gamma(A)\right|<\varepsilon$；

2) 对于任意的 $B'\subset B$，$\left|\gamma(B')-\gamma(A)\right|\ge\varepsilon$。

在这里，我们通过引入参数 ε 来扩大属性约简的范围，并尽可能减少冗余属性，同时将正负类样本间距离的变化保持在较小的范围内。例 1 更详细地描述了约简过程。

例 1：给出多标签数据集 $S=(U,A,L)$，如表 1 所示，样本集为 $U=\{x_1,x_2,x_3,x_4,x_5,x_6\}$，属性集为 $A=\{a,b,c\}$，标签集为 $L=\{l_1,l_2,l_3\}$。

Table 1. Multi-label data
表 1. 多标签数据

U	a	b	c	l_1	l_2	l_3
x_1	8	4	6	1	0	0
x_2	6	3	6	0	1	0
x_3	2	6	7	1	0	1
x_4	9	3	8	1	0	1
x_5	3	3	8	0	0	1
x_6	4	8	4	0	1	0

正负类样本集如表 2 所示。

Table 2. Sample set
表 2. 样本集

L	E	F
l_1	$E_1=\{x_1,x_3,x_4\}$	$F_1=\{x_2,x_5,x_6\}$
l_2	$E_2=\{x_2,x_6\}$	$F_2=\{x_1,x_3,x_4,x_5\}$
l_3	$E_3=\{x_3,x_4,x_5\}$	$F_3=\{x_1,x_2,x_6\}$

根据距离定义计算出正负类样本间距离，然后得到依赖度函数值，结果如表 3 所示。

Table 3. Distance
表 3. 距离

B	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	A
d_B	2	1	0	2.236	2	1	2.236
d_B	2	1	0	2.236	2	1	2.236
d_B	2	2	1	2.828	3.605	2.236	4.123
$\gamma(B)$	6	4	1	7.3	7.605	4.236	8.595

根据定义 3 计算得到表 4。

Table 4. Dependence degree
表 4. 依赖度

	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	A
$\lvert\gamma(B)-\gamma(A)\rvert$	2.595	4.595	7.595	1.295	0.99	4.359	0

根据定义 3 中的参数 $\varepsilon=1$，则从表 4 中可看出属性集 $\{a,c\}$ 是正负类边界距离的属性约简。

现在，我们设计了一种基于正负类边界距离的属性约简算法。把定义 3 中的参数 ε 看作算法的停止条件，当依赖度的增加小于参数 ε 时，算法终止，完成属性约简。

算法： 基于正负类边界距离的属性约简算法 PNBR

输入： (U,A,L)，参数 $\varepsilon>0$

输出： 约简 B

1： $A\to A_k$，$\phi\to B_k$

2： for $j=1,\cdots,|A_k|$

3： 对于每个属性 $a_j\in A_k$。计算出距离 $d_{B_k\cup\{a_j\}}(E_i,F_i)$

4： 计算出 $\gamma\left(B_k\cup\{a_j\}\right)$

5： end for

6： $j_0=\arg\max\limits_{j\le length(A_k)}\gamma\left(B_k\cup\{a_j\}\right)$

7： $A_k=A_k-\{a_{j_0}\},B_k=B_k\cup\{a_{j_0}\}$

8： 如果 $\left|\gamma\left(B_k\cup\{a_{j0}\}\right)-\gamma\left(B_k\right)\right|\le\varepsilon$

9： 输出属性约简 $B=B_k$

10： else

11： 回到第 2 步

12： end if

变量 A_k 是剩余的属性集，变量 B_k 是选择的属性集。最初的剩余属性集就是全部的属性集 A，从 A_k 中任意选择一个 a_j，把它并到 B_k 里，计算依赖度 $\gamma\left(B_k\cup\{a_j\}\right)$，有多少个属性就算多少个依赖度。然后选择依赖度最大的属性，在剩余属性 A_k 里去掉，选择的属性 B_k 中加上。选择最大的依赖度作为新的依赖度，当依赖增量小于参数 ε 时，算法终止，属性约简完成。当依赖增量大于参数 ε 时，回到第 2 步，选择了一个属性后还剩多少，然后循环上述过程，循环到满足不等式时，属性约简完成。在算法中，$|U|$ 是样本数，$|A|$ 是属性数，步骤 2 到步骤 5 是一个循环过程，循环 $|A|$ 次。因此，时间复杂度为 $O\left(|A||U|^2\right)$。

4. 实验

为了评估 PNBR 算法的性能，本文在九个多标签数据集上实现了 PNBR 算法，并将实验结果与四种多标签属性约简算法进行了性能比较。这些算法包括 PR、MDMR、FSSL 和 NLD，其中 PR 代表经典的正域约简算法[12]、MDMR 代表标签最大依赖和最小冗余算法[13]、FSSL 代表基于流标签的多标签学习的特征选择算法[11]、NLD 代表邻域标签依赖度约简算法[14]和 PNBR 代表我们提出的正负类边界距离的多标签数据属性约简算法。九个多标签数据集如表 5 所示。

Table 5. Multi-label data sets
表 5. 多标签数据集

Data set	Type	Samples	Attributes	Label
Flags	Hybrid	194	19	7
Yeast	Numerical	2417	103	14
Emotion	Numerical	593	72	6
CAL500	Numerical	500	62	174
Scene	Numerical	2407	294	6
Enron	Nominal	1702	1001	53
Medical	Nominal	978	1449	45
Core15k	Nominal	5000	499	374
Genbase	Nominal	662	1185	27

在我们的实验中，采用十折交叉验证来评估不同方法的有效性。十折交叉验证是将原始数据集随机分成 10 个大小相同的样本子集，轮流将其中 9 个子集当作多标签训练集，一个子集当作多标签测试集，进行实验。每次实验都会得出相应的正确率或差错率，10 次实验结果的正确率或差错率的平均值作为对算法精度的估计。表 6 给出了实验后所选属性的平均数量，并用下划线突显了两个最佳结果。通过比较所选属性的平均数量可知，MDMR 算法和 PNBR 算法更好，可以去除更多的冗余属性。而 PR 算法和 FSSL 算法相对差一点。

Table 6. Average numbers of the selected attributes
表 6. 所选属性的平均数量

Data set	Raw data	PR	MDMR	FSSL	NLD	PNBR
Flags	19	8.3	5.0	11.9	9.4	4.0
Yeast	103	7.7	7.3	9.6	9.4	8.6
Emotion	72	5.9	5.8	5.7	6.1	5.6
CAL500	62	8.7	8.5	8.2	3.7	3.3
Scene	294	8.9	8.6	11.9	8.8	11.8
Enron	1001	100.9	39.7	63.8	36.4	37.3
Medical	1449	63.9	34.4	63.3	31.9	31.2
Core15k	499	168.5	34.9	31.8	159.9	155.1
Genbase	1185	18.9	15.7	17.9	15.9	15.1

实验结果分析，对于数值型数据而言，PNBR 算法和 MDMR 算法选择的属性最少。在 Flags 数据集中，它只保留了 19 个属性中的 4 或 5 个，这两种算法相对其他算法较差。然而在数据集 Enron、Medical 和 Genbase 中，随着数据属性数量的增加，属性减少的就不多了，这种影响就可以忽略不计了。

十折交叉验证的平均运行时间如表 7 所示，可以看出，MDMR 算法、NLD 算法和 PNBR 算法的运行时间比其他算法长，PR 算法和 FSSL 算法的运行时间比其他算法短。

Table 7. Average running time
表 7. 平均运行时间

Data set	PR	MDMR	FSSL	NLD	PNBR
Flags	0.10	0.18	0.12	0.25	0.37
Yeast	107.34	255.62	329.41	757.11	967.71
Emotion	0.49	0.64	0.43	0.56	0.53
CAL500	0.37	3.96	1.52	5.31	9.38
Scene	320.68	318.62	848.60	841.88	821.58
Enron	337.68	373.22	208.66	337.81	324.85
Medical	193.33	129.91	150.05	92.38	88.69
Core15k	349.89	3768.82	264.80	2925.93	2815.86
Genbase	67.81	67.92	57.69	60.78	67.89

实验结果分析，对于数值型数据而言，FSSL 算法和 NLD 算法的计算时间大约是 PR 算法和 MDMR

算法的 2 倍，NLD 算法和 PNBR 算法的计算时间大约是 PR 算法和 MDMR 算法的 3 倍。在 Flags 数据集中，由于数据集较小，所以算法的计算速度的差异不明显。

本文用分类器 ML-9NN 评估了五种属性约简算法的分类性能，同时，采用 Hamming Loss、F_1 score 和 Coverage 三种多标签分类的评价指标，来衡量约简数据的分类精度[15]，其中下划线突显了两个最佳结果。

Table 8. Hamming loss
表 8. 汉明损失

Data set	Raw data	PR	MDMR	FSSL	NLD	PNBR
Flags	0.284 ± 0.051	0.285 ± 0.049	0.276 ± 0.057	0.269 ± 0.029	0.293 ± 0.031	0.289 ± 0.035
Yeast	0.198 ± 0.011	0.213 ± 0.010	0.202 ± 0.011	0.209 ± 0.012	0.199 ± 0.011	0.198 ± 0.010
Emotion	0.189 ± 0.019	0.238 ± 0.029	0.235 ± 0.029	0.198 ± 0.019	0.217 ± 0.019	0.179 ± 0.019
CAL500	0.167 ± 0.013	0.169 ± 0.011	0.165 ± 0.011	0.168 ± 0.013	0.159 ± 0.012	0.157 ± 0.013
Scene	0.118 ± 0.011	0.115 ± 0.010	0.123 ± 0.012	0.115 ± 0.011	0.131 ± 0.010	0.112 ± 0.011
Enron	0.065 ± 0.010	0.066 ± 0.011	0.057 ± 0.015	0.069 ± 0.011	0.059 ± 0.010	0.068 ± 0.012
Medical	0.029 ± 0.011	0.027 ± 0.010	0.025 ± 0.011	0.026 ± 0.013	0.019 ± 0.010	0.020 ± 0.011
Core15k	0.013 ± 0.000	0.014 ± 0.000	0.017 ± 0.000	0.015 ± 0.000	0.012 ± 0.000	0.011 ± 0.000
Genbase	0.019 ± 0.000	0.010 ± 0.000	0.019 ± 0.000	0.017 ± 0.000	0.021 ± 0.000	0.022 ± 0.000

汉明损失反应的是错误分类的样本标签占总样本标签的比率，该指标取值越小，算法的性能越好，当值为 0 时达到最优。从表 8 可以看出，NLD 算法和 PNBR 算法优于其他算法，MDMR 算法和 FSSL 算法的性能略优于 PR 算法，结合表 6 和表 7 中的属性约简数量和运行时间，NLD 算法和 PNBR 算法保留的属性更多，即使耗时长，但它的分类效果更好，特别是对于符号型数据。

Table 9. F_1 score
表 9. F_1 分数

Data set	Raw data	PR	MDMR	FSSL	NLD	PNBR
Flags	0.686 ± 0.059	0.695 ± 0.059	0.703 ± 0.079	0.687 ± 0.038	0.683 ± 0.049	0.679 ± 0.039
Yeast	0.627 ± 0.018	0.589 ± 0.011	0.597 ± 0.013	0.589 ± 0.018	0.625 ± 0.018	0.633 ± 0.011
Emotion	0.636 ± 0.047	0.627 ± 0.029	0.613 ± 0.026	0.631 ± 0.038	0.623 ± 0.057	0.615 ± 0.057
CAL500	0.448 ± 0.013	0.453 ± 0.013	0.446 ± 0.011	0.443 ± 0.011	0.435 ± 0.015	0.439 ± 0.016
Scene	0.669 ± 0.028	0.657 ± 0.039	0.661 ± 0.047	0.659 ± 0.028	0.657 ± 0.018	0.673 ± 0.015
Enron	0.447 ± 0.049	0.453 ± 0.058	0.451 ± 0.059	0.436 ± 0.088	0.439 ± 0.059	0.443 ± 0.059
Medical	0.367 ± 0.089	0.355 ± 0.117	0.359 ± 0.108	0.353 ± 0.117	0.373 ± 0.169	0.371 ± 0.168
Core15k	0.087 ± 0.018	0.083 ± 0.048	0.090 ± 0.017	0.082 ± 0.029	0.088 ± 0.017	0.081 ± 0.017
Genbase	0.873 ± 0.018	0.868 ± 0.018	0.869 ± 0.018	0.879 ± 0.018	0.878 ± 0.018	0.883 ± 0.018

F_1 score 相对于汉明损失是一个综合版本，F_1 score 衡量分类的准确性，并忽略错误分类的样本。它是精确率和召回率的调和平均值，该指标取值越大，算法的性能越好，通常最大值为 1，最小值为 0，当值为 1 时达到最优。从表 9 可知，除数据集 Core15k 外，五种算法在 F_1 score 方面没有显著差异。对于

Core15k 数据集，NLD 算法和 MDMR 算法的分类精度与原始数据差不多。

Table 10. Coverage
表 10. 覆盖率

Data set	Raw data	PR	MDMR	FSSL	NLD	PNBR
Flags	0.919 ± 0.019	0.927 ± 0.019	0.936 ± 0.028	0.923 ± 0.019	0.900 ± 0.010	0.899 ± 0.019
Yeast	0.869 ± 0.010	0.839 ± 0.010	0.840 ± 0.010	0.827 ± 0.010	0.836 ± 0.010	0.843 ± 0.010
Emotion	0.678 ± 0.027	0.543 ± 0.019	0.517 ± 0.019	0.658 ± 0.029	0.647 ± 0.03	0.639 ± 0.019
CAL500	0.987 ± 0.010	0.987 ± 0.010	0.989 ± 0.010	0.986 ± 0.010	0.975 ± 0.010	0.973 ± 0.010
Scene	0.634 ± 0.019	0.228 ± 0.010	0.243 ± 0.010	0.298 ± 0.010	0.339 ± 0.078	0.223 ± 0.010
Enron	0.717 ± 0.067	0.778 ± 0.100	0.789 ± 0.069	0.637 ± 0.078	0.725 ± 0.083	0.710 ± 0.063
Medical	0.418 ± 0.087	0.379 ± 0.087	0.419 ± 0.087	0.369 ± 0.119	0.423 ± 0.119	0.421 ± 0.119
Core15k	0.349 ± 0.039	0.336 ± 0.068	0.362 ± 0.059	0.351 ± 0.048	0.213 ± 0.029	0.227 ± 0.029
Genbase	0.953 ± 0.010	0.947 ± 0.019	0.964 ± 0.010	0.957 ± 0.028	0.953 ± 0.010	0.947 ± 0.010

覆盖率表示算法生成的排序底部的真实标签的排序平均值。排序越大越不相关，所以这个覆盖率越小，算法性能越好。从表 10 看出，PR 算法和 PNBR 算法优于其他算法。从数值上看，约简前后数据的分类精度并没有明显差异。实验结果表明所提约简算法能够建立合理的属性重要度排序，有效地去除冗余属性。

5. 总结

本文受基于集合到集合距离的启发，设计了一种基于正负类边界距离的属性约简算法，通过实验对所提的约简方法进行了评估，结果表明，所提约简算法能够建立合理的属性重要度排序，有效地去除冗余属性，并保持较高的分类精度。

实验研究表明，我们这个方法在处理大规模数据时，计算复杂度还是有点高，如何降低时间复杂度是我们今后要做的工作。另外在本文，我们将多标签数据分解中的标签看作为独立的，不相关的，但是在现实生活中，标签之间应该是有相关性的，所以，今后的研究工作中要将标签的相关性考虑进去，再做研究。

参考文献

[1] Zhang, M.L. and Zhou, Z.H. (2014) A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 1819-1837. https://doi.org/10.1109/TKDE.2013.39

[2] Xu, S.P., Yang, X.B., Yu, H.L., Yu, D.J., Yang, J.Y., Eric, C.C. and Tsang, J. (2016) Multi-Label Learning with Label-Specific Feature Reduction. *Knowledge-Based Systems*, **104**, 52-61. https://doi.org/10.1016/j.knosys.2016.04.012

[3] Zhu, P.F., Xu, Q., Hu, Q.H., Zhang, C.Q. and Zhao, H. (2018) Multi-Label Feature Selection with Missing Labels. *Pattern Recognition*, **74**, 488-502. https://doi.org/10.1016/j.patcog.2017.09.036

[4] Li, H., Li, D.Y., Zhai, Y.H., Wang, S.G. and Zhang, J. (2016) A Novel Attribute Reduction Approach for Multi-Label Data Based on Rough Set Theory. *Information Sciences*, **367-368**, 827-847. https://doi.org/10.1016/j.ins.2016.07.008

[5] Li, Y.W., Lin, Y.J., Liu, J.H., Weng, W., Shi, Z.K. and Wu, S.X. (2018) Feature Selection for Multi-Label Learning Based on Kernelized Fuzzy Rough Sets. *Neurocomputing*, **318**, 271-286. https://doi.org/10.1016/j.neucom.2018.08.065

[6] Liu, J.H., Lin, Y.J., Li, Y.W., Weng, W. and Wu, S.X. (2018) Online Multi-Label Streaming Feature Selection Based on Neighborhood Rough Set. *Pattern Recognition*, **84**, 273-287. https://doi.org/10.1016/j.patcog.2018.07.021

[7] Lin, Y.J., Hu, Q.H., Zhang, J. and Wu, X.D. (2016) Multi-Label Feature Selection with Streaming Labels. *Information*

Sciences, **372**, 256-275. https://doi.org/10.1016/j.ins.2016.08.039

[8] Fan, X.D., Zhao, W.D., Wang, C.Z. and Huang, Y. (2018) Attribute Reduction Based on Max-Decision Neighborhood Rough Set Model. *Knowledge-Based Systems*, **151**, 16-23. https://doi.org/10.1016/j.knosys.2018.03.015

[9] Lin, Y.J., Li, Y.W., Wang, C.X. and Chen, J.K. (2018) Attribute Reduction for Multi-Label Learning with Fuzzy Rough Set. *Knowledge-Based Systems*, **152**, 51-61. https://doi.org/10.1016/j.knosys.2018.04.004

[10] Wang, C.X., Lin, Y.J. and Liu, J.H. (2019) Feature Selection for Multi-Label Learning with Missing Labels. *Applied Intelligence*, **49**, 3027-3042. https://doi.org/10.1007/s10489-019-01431-6

[11] Liu, J.H., Li, Y.W., Weng, W., Zhang, J., Chen, B.H. and Wu, S.X. (2020) Feature Selection for Multi-Label Learning with Streaming Label. *Neurocomputing*, **387**, 268-278. https://doi.org/10.1016/j.neucom.2020.01.005

[12] 段洁, 胡清华, 张灵均, 钱宇华, 李德玉. 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.

[13] Lin, Y.J., Hu, Q.H., Liu, J.H. and Duan, J. (2015) Multi-Label Feature Selection Based on Max-Dependency and Min-Redundancy. *Neurocomputing*, **168**, 92-103. https://doi.org/10.1016/j.neucom.2015.06.010

[14] Fan, X.D., Chen, Q., Qiao, Z.J., Wang, C.Z. and Ten, M.Y. (2020) Attribute Reduction for Multi-Label Classification Based on Labels of Positive Region. *Soft Computing*, 1-11. https://doi.org/10.1007/s00500-020-04780-4

[15] Zhang, M.L. and Zhou, Z.H. (2006) ML-KNN: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recognition*, **40**, 2038-2048. https://doi.org/10.1016/j.patcog.2006.12.019