Hans Journal of Data Mining 数据挖掘, 2020, 10(4), 240-246
Published Online October 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.104025

Hans汉斯

基于链分解的多标签分类属性约简

张　莹

渤海大学数学系，辽宁　锦州
Email: 979542238@qq.com

收稿日期：2020年9月4日；录用日期：2020年9月18日；发布日期：2020年9月27日

摘　要

本文提出了基于链分解的多标签属性约简方法。通过考虑标签之间的相关性，将标签进行排序，根据排序方法，多标签问题被分解成单标签链的形式，对于链中每一个子问题通过粗糙集方法重新定义下近似、正域、依赖度，并进行属性约简。实验结果表明，该方法能在不降低分类精度的情况下去除大部分冗余属性。

关键词

多标签分类，属性约简，粗糙集，链分解

Attribute Reduction for Multi-Label Classification Based on Chain Decomposition

Ying Zhang

Department of Mathematics, Bohai University, Jinzhou Liaoning
Email: 979542238@qq.com

Received: Sep. 4th, 2020; accepted: Sep. 18th, 2020; published: Sep. 27th, 2020

Abstract

In this paper, a new multi-label attribute reduction algorithm based on the chain decomposition is proposed. Considering the correlation between the labels, the labels are sorted. According to the sorting method, the multi-label problem is decomposed into a single-label problem chain. For each sub-problem, the lower approximation, the positive region and the dependency are redefined by the rough set method, and the attributes are reduced. Experimental results show that the algorithm can remove most of the redundant attributes without reducing the classification accuracy.

文章引用: 张莹. 基于链分解的多标签分类属性约简[J]. 数据挖掘, 2020, 10(4): 240-246.
DOI: 10.12677/hjdm.2020.104025

Keywords

Attribute Reduction, Rough Set, Multi-Label Classification, Chain Decomposition

1. 引言

在传统监督学习中一个样本只与一个标签相关，这类问题被称为单标签问题。但是在现实生活中往往并非如此，一个样本也可以与多个标签相关联，比如一篇文章可能存在多个关键词，一幅图像可以拥有多个主题，我们把这类问题称为多标签问题。与单标签分类不同，多标签分类问题会更加复杂。

问题转换是处理多标签问题的方法之一。其主要思想是将多标签问题转化为一个或多个单标签问题进行处理。BR (binary relevance)是最常见的问题转换方法，实现方法简单，容易理解。但在考虑标签之间的相关性时，最终构建模型的泛化能力会比较弱。而 Read 等[1]在 2009 年提出的分类器链算法，在一定程度上克服了这个问题。分类器链同样是将多标签问题转化为单标签问题[2]，但与传统二分类方法不同的是，分类器链算法把标签当作额外信息添加到属性集中，即每个已知标签都可以看作是属性空间的子集。实际上就是样本属性在不断的扩充。在这一过程中考虑了标签之间的相关性，特别是在训练样本很少的情况，缺少有用的信息时，考虑标签之间的相关性就显得尤为重要。

粗糙集是一种新的软计算方法，近年来受到越来越多的关注。它的有效性已经在许多科学和工程领域的成功应用中得到了证明。最早由波兰科学家 Pawlak 在 1982 年[3]提出。此后，粗糙集理论逐渐应用于单标签数据的属性约简中[4]，并取得了令人满意的效果。近年来，粗糙集被广泛地应用于多标签数据属性约简中[5] [6] [7] [8]。然而，在约简过程中考虑标签之间的相关性，降低计算复杂度是需要解决的主要问题。本文主要根据多标签链分解的特点，将其与粗糙集方法相结合，在考虑标签间相关性的基础上进行属性约简。

本文剩余部分结构如下：在第二节中，提出了两种标签排序方法，并将多标签分解成链的形式。在第三节中，对于每个分解之后的子问题给出了新的相似类、正域、依赖度的定义，并设计了一种新的属性约简算法。在第四节中，在给定的五个数据集上进行了数值实验，并对于实验结果进行了分析。在第五节中，对本文所得的结论和实验结果进行总结。

2. 多标签链分解

基于链分解的多标签问题本质上是将多标签问题转化为链的形式。在分解过程中，已知标签依次作为额外的属性为样本提供分类信息，所以标签的排序非常重要。本节主要提出两种标签排序方法，建立了链式分解。在此之前给出多标签分类问题的基本框架。

令 $X=\{x_1,x_2,x_3,\cdots,x_n\}$ 为样本集， $Y=\{y_1',y_2',y_3',\cdots,y_m'\}$ 为标签集， $A=\{a_1,a_2,a_3,\cdots,a_d\}$ 代表属性集合。我们可以将多标签数据集表示为 (X,A,Y)。对每一个属性 $a\in A$，样本 x_i 在属性 a 上的取值记为 $a(x_i)$。对每个标签 $y_j'\in Y$，样本 x_i 的标签值为 $y_j'(x_i)$，如果 x_i 具有标签 y_j'， $y_j'(x_i)=1$ 否则为 0。下面我们给出两种标签排序的方法。

方法一：邻域法

该方法主要是利用邻域的概念，找出每个样本邻域内各个标签的出现次数，对于所有样本邻域内同一类标签出现次数相加并求其平均值，根据平均值确定标签排列顺序。接下来我们将给出邻域法的相关定义。

给定标签数据集(X,A,Y)，对于任意$\varepsilon>0$，样本x_i的相似类定义如下：

$$[x_i]_A^{\varepsilon}=\{x\in X:|a(x_i)-a(x)|\le\varepsilon,a\in A\}$$

对于任意$y_j'\in Y$，令

$$T_j(x_i)=\sum_{x\in[x_i]_A^{\varepsilon}}y_j'(x),\ j=1,2,\cdots,m,\ i=1,2,\cdots,n.$$

其中x是x_i邻域内的任意一个样本，$T_j(x_i)$表示样本x_i邻域内标签y_j'出现次数。基于以上公式，将所有样本的邻域内标签y_j'出现的次数相加并求其平均值：

$$rank(y_j')=\frac{\sum_{i=1}^{n}T_j(x_i)}{n}$$

根据平均数的大小，按照降序将标签进行排列。不妨设排序后的标签列为$(y_1,y_2,\cdots,y_m)$。

方法二：计数法

计数法就是找到所有与标签y_j'相关的样本，根据相关样本集基数的大小进行标签排序。接下来我们将给出相关定义。

给定标签数据集(X,A,Y)，令

$$X_j=\{x_i\mid y_j'(x_i)=1,1\le i\le n\},\ j=1,2,\cdots,m.$$

这里，$X_j\subset X$是与标签y_j'相关的样本集，根据X_j基数将标签降序排列，排序后的标签列仍然记为$(y_1,y_2,\cdots,y_m)$。

根据以上两种标签排序方法，对m个标签进行排序。将标签依次视为新的属性加入到原属性集中，令

$$A_1=A\cup\{y_1\}$$

对于序列中的第j个标签新的属性集可以表示为：

$$A_j=A_{j-1}\cup\{y_j\},\ j=2,3,\cdots,m.$$

则有$A\subset A_1\subset A_2\subset\cdots\subset A_{j-1}\subset A_m$。

以上是将多标签问题分解为链式子问题的过程，其主要优势是考虑了标签之间的相关性，标签被当做属性添加到原始属性集A中用来为样本提供有效信息。此时，子问题的数据集记为$(X,A_j,y_j),j=1,2,\cdots,m$。

3. 属性约简

在本节中根据以上多标签链分解的特点，对于每个子问题将重新定义下近似、正域并提出新的约简方法，在此之前我们将先给出标签信息集的定义。对于任意$y_j\in Y$，标签信息集定义如下：

$$E_j=\{x\in X:y_j(x)=1\},\ j=1,2,\cdots,m.$$

由所有标签信息集组成的集合族可以表示为：

$$E=\{E_j\mid j=1,2,\cdots,m\}.$$

对于任意属性子集$B\subset A$，根据属性集A_j的特点，扩充后的属性集可以表示成如下形式：

$$B_1 = B \cup \{y_1\}$$

$$B_j = B_{j-1} \cup \{y_j\},\ j = 1,2,\cdots,m.$$

定义 2.1：对于多标签数据集(X, A_j, Y)，样本 x_i 的相似类可以被重新定义为。

$$[x_i]_{B_j}^{\varepsilon} = \{x \in X : |a(x_i) - a(x)| \le \varepsilon, |y_j(x_i) - y_j(x)| \le \varepsilon, a, y_j \in B_j\}.$$

定义 2.2：给定多标签数据集(X, A_j, Y)，对于属性子集 $B_j \subset A_j$，关于标签信息集 E_j 的下近似定义如下：

$$\underline{R}_{B_j}^{\varepsilon}(E_j) = \{x_i \in X : [x_i]_{B_j}^{\varepsilon} \subseteq E_j\}.$$

定义 2.3：给定原始多标签数据集(X, A, Y)，对于属性子集 $B \subset A$，其正域定义为：

$$POS_B^{\varepsilon}(E) = \bigcup_{E_j \in E} \underline{R}_{B_j}^{\varepsilon}(E_j).$$

引理 2.1：对于任意的属性子集 $B \subset A, B' \subset A$，如果 $B' \subset B$，则有

$$POS_{B'}^{\varepsilon}(E) \subseteq POS_B^{\varepsilon}(E)$$

证明：对于任意 $x \in POS_{B'}^{\varepsilon}(E)$ 从定义 2.3 可知 $x \in \bigcup_{E_j \in E} \underline{R}_{B_1}^{\varepsilon}(E_j)$。故存在 $E_j \in E$，使得 $x \in \underline{R}_{B'}^{\varepsilon}(E_j)$ 根据下近似定义 $[x]_{B'}^{\varepsilon} \subseteq E_j$，由 $B' \subset B$ 可以得到 $[x]_B^{\varepsilon} \subseteq [x]_{B'}^{\varepsilon} \subseteq E_j$，因此 $x \in \underline{R}_B^{\varepsilon}(E_j)$，则有 $x \in \bigcup_{E_j \in E} \underline{R}_B^{\varepsilon}(E_j)$，即 $x \in POS_B^{\varepsilon}(E)$，所以 $POS_{B'}^{\varepsilon}(E) \subseteq POS_B^{\varepsilon}(E)$ 成立。

定义 2.4：对于多标签数据(X, A, Y)，当且仅当 $B \subseteq A$ 满足以下条件

1) $POS_B^{\varepsilon}(E) = POS_A^{\varepsilon}(E)$，

2) $POS_{B'}^{\varepsilon}(E) \neq POS_A^{\varepsilon}(E)$ 其中 $B' \subset B$，

则称集合 B 是多标签数据集的链式正域约简。

接下来，我们将介绍链式依赖性约简。在此之前，将依赖函数定义为：

$$\gamma_B^{\varepsilon}(E) = \frac{1}{m} \sum_{j=1}^{m} \frac{\left|\underline{R}_B^{\varepsilon}(E_j)\right|}{|X|}$$

其中 $|\cdot|$ 表示相应集合的基数。由引理 2.1 可以直接得到依赖度的单调性引理。

引理 2.2：对于任意的 $B' \subset B \subset A$，有

$$\gamma_{B'}^{\varepsilon}(E) \le \gamma_B^{\varepsilon}(E)$$

定义 2.4：对于多标签数据(X, A, Y)，对于任意的 $B' \subset B$ 如果属性子集 $B \subseteq A$ 满足以下条件：

1) $\gamma_B^{\varepsilon}(E) = \gamma_A^{\varepsilon}(E)$，

2) 对于任意的 $B' \subset B$，$\gamma_{B'}^{\varepsilon}(E) \neq \gamma_A^{\varepsilon}(E)$，

则称集合 B 是多标签数据集的链式依赖度约简。

在样本空间中，每个样本包含多个属性，而且每个属性的重要度不同。接下来，我们将使用上面定义的依赖函数来评估每个属性的重要性。

定义 2.5：给定多标签数据(X, A, Y)，令属性子集 $B \subseteq A$，则属性 $a_i \in B$ 关于属性子集 B 的重要性度定义为：

$$Sig_{in}(a_i,B)=\gamma_B^{\varepsilon}(E)-\gamma_{(B-a_i)}^{\varepsilon}(E).$$

这里我们使用重要性的概念来给出算法的伪代码。

算法：链式依赖度约简
输入：(X,A,Y)，终止参数 $\varphi>0$ 。
输出：约简 Red
1：$Red=\varnothing,B=A$
2：for $i=1:lenght\|B\|$
3：对于每个属性计算重要度 $sig_{in}(a_i,B)$
4：end for
5：选择重要度最大的属性 $sig_{in}(a_k,B)=\max\{a_i\in B:sig_{in}(a_i,B)\}$
6：$Red=Red\cup\{a_k\}$
7：$B=B-\{a_k\}$
8：If $sig_{in}(a_k,B)\le\varphi$
9：　　输出约简属性
10：else
11：　　返回到第 2 步
12：end if

参数 φ 用来构造终止准则，当待选属性的重要度小于 φ 时，算法停止运行并输出约简属性。

第 2 步到第 6 步是计算属性集 B 中各属性的重要度，根据定义 2.5，将值最大的属性添加到集合 Red。第 7 步到第 11 步，判断算法的终止条件，当待选属性的重要度小于 φ 时输出约简，否则返回第 2 步。

4. 实验结果

为了评估本文约简算法的性能，选择 5 个多标签数据集。在每一个数据集上与其他 3 种算法进行对比，如表 1，其中 PRR 代表正域约简、MLFRS 代表多标签模糊粗糙集属性约简[9]、NLDR 代表邻域标签依赖度约简[10]、CDDR 是本文提出的链式依赖度约简，将不同算法的约简数据输入到分类器中。为了比较约简效果，对每个数据集采用统一的约简比，即不同算法约简的数据包含相同数量的属性。同时，计算未约简数据的度量作为参考。根据样本个数将数据集随机分成 10 等份，每部分 80%作为训练集，20%作为测试集，取 10 次独立实验的平均值作为实验的最终结果。

Table 1. Multi-label data sets
表 1. 多标签数据集

Date set	Type	Samples	Attributes	Label
Emotion	Numerical	593	72	6
CAL500	Numerical	500	62	174
Yeast	Numerical	2417	103	14
Medical	Nominal	978	1449	45
Genbase	Nominal	662	1185	27

Hamming Loss 表示测试样本中被误分类的标签在样本所有标签中占的比例。值越小分类能力越强。显然，评价指标与数据集中原始标签的数量密切相关，当数据集中标签数量增加时，其值可能会增加。因此，对于不同的数据集，它是不可比较的。表 2 的第 2 列给出了原始数据的 Hamming Loss 值。第 3 至第 6 列中分别是五种算法的简化数据值。从表中可以看出，该算法在数据集 CAL500、Yeast、Genbase

上的性能明显优于其他算法。而在数据集 Medical 上与最优算法也仅相差 0.004。

Table 2. Hamming loss
表 2. 汉明损失

Date set	Raw data	PRR	MLFRS	NLDR	CDDR
Emotion	0.267	0.289	**0.272**	0.291	0.290
CAL500	0.110	0.138	0.142	0.133	**0.130**
Yeast	0.213	0.269	0.264	0.292	**0.229**
Medical	0.011	0.022	**0.014**	0.019	0.018
Genbase	0.044	0.045	0.047	0.050	**0.043**

Table 3. Coverage
表 3. 覆盖率

Date set	Raw data	PRR	MLFRS	NLDR	CDDR
Emotion	2.305	2.206	2.242	**2.202**	0.207
CAL500	113.1	115.7	132.8	133.7	**114.4**
Yeast	6.112	7.109	7.168	7.192	**6.129**
Medical	2.491	3.512	3.478	**3.462**	3.486
Genbase	0.562	0.656	0.559	0.568	**0.556**

Coverage 用于考察在样本的类别标签排序序列中，覆盖所有相关标签所需要的搜索深度情况，该指标取值越小性能越优。表 3 对于 Coverage 在给定的 5 个数据集中取得较好的效果，特别是在数据集 Yeast 上算法 CDDR 显示出较大的优势，同时在其他 2 个数据上 CDDR 的性能与最优算法性能也是非常接近的，没有太多差异。

根据表 2、表 3 可以看出，CDDR 与其他多标签属性约简算法相比具有很强的竞争力，其性能优势更明显地验证了基于 CDDR 的属性约简的有效性。

5. 总结

本文主要介绍了一种新的基于链分解的多标签约简方法。首先针对不同标签所提供信息的重要程度不同，给出了两种标签排序方法固定标签序列以避免错误信息的传递。其次根据固定的序列将多标签问题分解成链的形式，将分解的链与模糊方法相结合，对于每个子问题重新给出定义。最后在不影响精度的基础上，除去冗余属性进行约简。由实验范围广泛的数据集表明，我们的方法具有高度可比性。

值得注意的是，约简本身是一个不连续优化问题，很多算法不能应用。而 CDDR 算法又只适用于小型数据集，存在计算复杂度高的问题。对于标签之间的关系本文主要从标签出现次数上进行考虑其相关性，但是标签之间的固有联系是一个比较复杂的问题。未来，我们将继续优化模型，寻求更好的排序方法。同时将更深入地研究如何解决标签间相关性问题。

参考文献

[1] Read, J., Pfahringer, B., Holmes, G. and Frank, E. (2009) Classifier Chains for Multi-Label Classification. In: Buntine, W., Grobelnik, M. and Shawe-Taylor, J., Eds., *Lecture Notes in Artificial Intelligence* 5782, Springer, Berlin, 254-269. https://doi.org/10.1007/978-3-642-04174-7_17

[2] Read, J., Pfahringer, B., Holmes, G. and Frank, E. (2011) Classifier Chains for Multi-Label Classification. *Machine.*

Learning, **85**, 333-359. https://doi.org/10.1007/s10994-011-5256-5

[3] Pawlak, Z. (2011) Rough Sets. *International Journal of Computer and Information Sciences*, **11**, 34l-356. https://doi.org/10.1007/BF01001956

[4] Hu, Q.H., Yu, D.R., Liu, J.F. and Wu, C. (2008) Neighborhood Rough Set Based Heterogeneous Feature Subset Selection. *Information Sciences*, **178**, 3577-3594. https://doi.org/10.1016/j.ins.2008.05.024

[5] Li, H., Li, D., Zhai, Y., Wang, S. 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

[6] Lin, Y., Li, Y., Wang, C. and Chen, J. (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

[7] Liu, J., Lin, Y., Li, Y., Weng, W. and Wu, S. (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

[8] Lin, Y., Hua, Q., Liu, J., Chen, J. and Duan, J. (2016) Multi-Label Feature Selection Based on Neighborhood Mutual Information. *Applied Soft Computing*, **38**, 244-256. https://doi.org/10.1016/j.asoc.2015.10.009

[9] Lin, Y., Li, Y., Wang, C. and Chen, J. (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] Fan, X., Chen, Q., Qiao, Z., Wang, C. and Ten, M. (2020) Attribute Reduction for Multi-Label Classification Based on Labels of Positive Region. *Soft Computing*, **24**, 14039-14049. https://doi.org/10.1007/s00500-020-04780-4