DOI:10.3969/j.issn.1673-4785.201209056
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20130409.1436.005.html

综合属性选择和删除的属性约简方法

杨成东[1], 邓廷权[2]
(1.临沂大学 信息学院, 山东 临沂 276005; 2.哈尔滨工程大学 理学院, 黑龙江 哈尔滨 150001)

摘 要:属性约简能有效地消除信息冗余,广泛应用于人工智能、机器学习.通过实例指出基于辨识矩阵的经典的属性约简方法存在不能得到约简的可能性,仍具有冗余性.因此,提出了综合属性选择和删除算法的辨识矩阵属性约简方法,并有效解决该问题.通过UCI标准数据集验证表明,新方法比经典方法进一步减少了属性的个数,凸显其实用性和有效性.

关键词:辨识矩阵;属性约简;信息冗余;人工智能;机器学习;属性选择;属性删除
中图分类号: TP301.6 **文献标志码**:A **文章编号**:1673-4785(2013)02-0183-04

An approach to attribute reduction combining attribute selection and deletion

YANG Chengdong[1], DENG Tingquan[2]
(1. School of Informatics, Linyi University, Linyi 276005, China; 2. College of Science, Harbin Engineering University, Harbin 150001, China)

Abstract: Attribute reduction has been defined as a method for removing information redundancy effectively, which has been widely applied to artificial intelligence, and machine learning. However, an example demonstrates classical attribute reduction approaches based on discernibility matrix may not get a reduction with redundancy. Therefore, an attribute reduction based on discernibility matrix combining attribute selection and deletion was proposed and thus, the problem was solved effectively. Moreover, UCI standard data sets provide further explanations on the feasibility, effectiveness, and as well as additional information on reducing the number of attributes without the classical approaches.

Keywords: discernibility matrix; attribute reduction; information redundancy; artificial intelligence; machine learning; attribute selection; attribute deletion

属性约简利用粗糙集[1-2]等理论,旨在保持信息系统决策能力不变的条件下,去除冗余属性,从而减少数据的冗余度,是机器学习和人工智能最重要的研究方向之一.属性约简方法有很多,譬如基于依赖度的属性约简方法[3]、基于互信息的属性约简方法[4-5]、基于模糊粗糙集的属性约简方法[6-8]等.Skowron于1992年提出了辨识矩阵和辨识函数的概念[9],利用辨识矩阵和辨识函数实现了属性约简,并得到了广泛的研究[10].然而,基于辨识矩阵的属性约简方法,存在不能得到约简的可能性,仍具有冗余性.

收稿日期:2012-09-25. **网络出版日期**:2013-04-09.
基金项目:山东省高等学校科技计划资助项目(J12LN91);山东省信息化与工业化融合专项课题资助项目(2012EI100).
通信作者:杨成东. E-mail: yangchengdong@lyu.edu.cn.

1 基础知识

给定决策系统 $S=(U,C\cap D,V,f)$,其辨识矩阵定义为

$$\boldsymbol{M}=\boldsymbol{M}(x,y),$$

式中:$\boldsymbol{M}(x,y)$ 定义为

$$\boldsymbol{M}(x,y)=\begin{cases}a\in C\mid f(x,a)\neq f(y,a),f(x,D)\neq f(y,D);\\ \varnothing,\text{其他}.\end{cases}$$

显然,矩阵 $\boldsymbol{M}$ 中元素 $\boldsymbol{M}(x,y)$ 是由处于不同决策类中的对象 x 和 y 属性值不同的属性组成.

辨识函数$f(\boldsymbol{M})$定义为

$$f(\boldsymbol{M}) = \wedge\{\vee \boldsymbol{M}(x,y) \mid \forall x,y \in U, \boldsymbol{M}(x,y) \neq \varnothing\}.$$

式中:$\vee$和$\wedge$是2个基本的二值逻辑运算:析取和合取. 辨识函数是一个布尔表达式,通过等价的逻辑计算,将其化成若干个小合取式的析取式,那么每个小合取式就是一个约简. 一般地,约简不是惟一的,决策系统的所有约简用$\mathrm{RED}_C(D)$表示.

辨识矩阵和辨识函数有如下性质:

1) 核是辨识矩阵中所有单个元素组成的集合;

2) 辨识函数$f(\boldsymbol{M})$的极小析取范式中的所有合取式是属性集C的所有约简.

辨识矩阵和辨识函数方法能够求出所有约简,因此具有十分重要的理论意义. 然而利用该方法求出的所有约简仍是一个NP-hard问题,特别是在大规模数据集中几乎无法求出约简,其速度非常慢,而实际中通常只需要一个约简.

2 辨识矩阵属性约简方法及其缺点

作为经典辨识矩阵算法,基于属性频率的辨识矩阵快速属性约简算法利用频率作为衡量属性重要程度的依据,具有重要的实用价值. 在辨识矩阵中出现频率最高的属性是较为重要的,优先选择该属性. 基于辨识矩阵属性频率的快速属性约简算法如下:

算法1 基于辨识矩阵属性频率的属性约简算法:

Input: $S=(U,C\cup D,V,f)$

Output: red

1) compute discernibility matrix $\boldsymbol{M}$.

2) if $\boldsymbol{M}=\boldsymbol{0}$

3) return red

4) end if

5) for $a \in C-\mathrm{red}$

6) compute the attribute frequency of a

7) end for

8) select a_k with the largest attribute frequency, then $\mathrm{red}=\mathrm{red}\cup\{a_k\}$;

9) set the elements of $\boldsymbol{M}$ including a_k with 0.

10) if $\boldsymbol{M}=0$.

11) return red;

12) else

13) turn to 5);

14) return.

该算法中的时间复杂度分为关键2步:一是对属性进行选择有2个循环,时间复杂度为$O(|C|^2)$;另一个是计算单个属性的频率,时间复杂度为$O(|U|)$. 因此总的时间复杂度为:$O(|U||C|^2)$.

例1 给定关于大豆质量的决策系统$S=(U,C\cup D,V,f)$如表1,其中$C=\{a,b,c,d,e\}$是条件属性,D是决策属性.

表1 决策系统

Table 1 A decision system

U	a	b	c	d	e	D
1	Middle	Middle	Middle	fer1	pes1	High
2	Middle	High	High	fer2	pes2	High
3	High	Middle	Middle	fer2	pes1	High
4	High	High	High	fer2	pes1	High
5	High	Middle	High	fer2	pes1	Low
6	High	High	Middle	fer1	pes1	Low
7	High	Middle	High	fer1	pes1	Low
8	High	High	Middle	fer1	pes1	Low

通过计算,该信息系统有2个约简,$\mathrm{RED}_C(D)=\{\{b,c\}\}$. 可以验证该系统是协调决策系统,见表1. 而利用算法1,求得的结果是$\{b,d,c\}$,显然$\{b,d,c\}\notin \mathrm{RED}_C(D)$,仍包含了冗余属性$\{d\}$. 该例说明经典算法不能有效计算约简,仍具有一定的冗余性. 本文提出一种新的属性约简方法来解决该问题.

3 结合属性选择和删除的属性约简方法

首先证明算法1得到的属性约简没有损失信息,即其依赖度相同.

定理1 给定决策系统$S=(U,C\cup D,V,f)$,经过算法1后,得到red,那么

$$\gamma_{\mathrm{red}}(D)=\gamma_C(D).$$

证明 反证法. 假设$\gamma_{\mathrm{red}}(D)\neq\gamma_C(D)$,那么,$\gamma_{\mathrm{red}}(D)<\gamma_C(D)$,因此,存在$x\in \mathrm{Pos}_C(D)$,使得$x\notin \mathrm{Pos}_{\mathrm{red}}(D)$,那么,存在$y\in[x]_{\mathrm{red}}$,使得$\boldsymbol{M}(x,y)\neq\varnothing$. 然而这与算法1矛盾,因为经过算法1运算后,$\boldsymbol{M}$是一个空矩阵,因此假设不成立.

例1说明了经典算法得到的red还具有一定冗余性,而定理1说明了经典算法得到的red与原始决策系统具有相同的分辨能力. 因此,本文提出能有效避免冗余的辨识矩阵属性约简快速算法.

算法2 结合属性选择和删除的属性约简快速算法:

Input: $S=(U,C\cup D,V,f)$

Output: red

1) compute discernibility matrix $\boldsymbol{M}$ according to red.

2) if $\boldsymbol{M}=0$

3) return red

4) end if

5) for $a\in C-\mathrm{red}$

6) compute the attribute frequency of a

7) end for

8) select a_k with the largest attribute frequency, then red = red $\cup a_k$;

9) set the elements of $\boldsymbol{M}$ including a_k with 0.

10) if $\boldsymbol{M}=0$,

11) return red;

12 else

13) turn to 5);

14) end if

15) for each $a \in$ red ,

16) if $\gamma_{\text{red}-a}(D)=\gamma_C(D)$,

17) red = red $-a$;

18) end if

19) end for

20) return red

算法 2 比算法 1 多了一个循环 $O(|U||C|)$,由于这 2 个循环是并列的,那么总的时间复杂度为 $O(|U||C|^2)+O(|U||C|)=O(|U||C|^2)$,因此算法 2 与算法 1 具有相同的时间复杂度,本文提出的算法的时间复杂度不会增加.

下面证明算法 2 选择的属性子集是约简,既保持了信息,又有效地消除了冗余信息.

定理 2　给定决策系统 $S=(U,C\cup D,V,f)$,经过算法 2 后,得到 red,那么 red 是约简.

证明　类似于定理 1 的证明,可以得到

$$\gamma_{\text{red}}(D) = \gamma_C(D).$$

另一方面,采用反证法证明 red 是独立的. 假设 red 不是独立的,那么存在 $a \in$ red,满足

$$\gamma_{\text{red}}(D) = \gamma_{\text{red}-a}(D).$$

那么这样的属性 a 在 14) ~ 19) 的循环中被删除了,即 $a \notin$ red.

因此,假设不成立,red 是独立的. 所以,red 是约简.

例 2　继续使用例 1,利用算法 2,可以得到约简 red = $\{b,c\}$. 因此,与基于辨识矩阵属性约简方法相比,该方法能够有效地获得约简.

4　实验与分析

用 6 个 UCI 标准数据集来验证本文提出的方法的实用性和有效性,如表 2 所示. 表 3 对经典方法选择的属性序列和本文提出的方法选择的属性序列进行了比较. 利用经典方法得到的 6 个数据集中,Heart、Lymph、Soybean 有 3 个不是约简. 而本文方法得到的都是约简,有效地解决了经典方法得不到约简的问题.

表 2　UCI 标准数据集

Table 2　UCI standard datasets

数据集	简称	对象个数	属性个数
Car Evaluation	Car	172	7
Spect Heart	Heart	267	23
Tic-Tac-Toe	Tic	958	10
Lymphography	Lymphography	148	19
Soybean (Large)	Soybean	683	36
Zoo	Zoo	101	17

表 3　与经典方法的比较

Table 3　Comparison of UCI standard datasets and the classical approaches

数据集	原始属性个数	经典方法		本文提出的方法	
		选择序列	选择个数	选择序列	选择个数
Car	7	{2,1,4,6,5,3}	6	{2,1,4,6,5,3}	6
Heart	23	{16,22,21,20,19,1,13,3,8,5,9,10,14,4,12 }	15	{16,22,21,20,19,1,13,3,5,9,14,12}	12
Lymph	10	{18,14,13,12,1,15,2,10}	8	{18,14,13,12,15,2,10}	7
Soybean	19	{6,5,7,9,16,4,10,1,22,29,15,28,13,21,14,8}	16	{6,5,7,9,16,4,10,1,22,29,15,8}	12
Tic	36	{5,2,4,6,8,1,3,7}	8	{5,2,4,6,8,1,3,7}	8
Zoo	17	{6,13,4,8,3}	6	{6,13,4,8,3}	6
平均属性个数	18.7	-	9.8	-	8.5

从选择属性个数来看,与原始数据集对比,经典算法和本文提出的方法都大大减少平均属性个数.进一步地,本文算法的平均属性个数为8.5,比经典算法减少了1.3个.因此,本文提出的方法能够获得更为精简的集约数据集,进一步降低了数据集的冗余性.

5 结束语

本文提出了结合属性选择和删除的属性约简方法,该方法能够彻底解决经典算法产生的冗余,得到有效的约简,解决了经典算法不能得到约简的问题.通过6个UCI标准数据集,实例分析表明,提出的方法选择的平均属性个数比经典算法减少了1.3个,显示了其有效性和实用性.

参考文献:

[1]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: 5-7.

[2]PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[3]蒋云良, 杨章显, 刘勇. 不协调信息系统快速属性分布约简方法[J]. 自动化学报, 2012, 38(3): 382-388.
JIANG Yunliang, YANG Zhangxian, LIU Yong. Quick distribution reduction algorithm in inconsistent information system[J]. Acta Automatica Sinica, 2012, 38(3): 382-388.

[4]XU F, MIAO D, WEI L. Fuzzy-rough attribute via mutual information with an application to cancer classification[J]. Computers and Mathematics with Applications, 2009, 57(6): 1010-1017.

[5]BHATT R, GOPAL M. On fuzzy-rough sets approach to feature selection[J]. Pattern Recognition Letters, 2004, 26(7): 965-975.

[6]胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640-649.
HU Qinghua, YU Daren, XIE Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. Journal of Software, 2008, 19(3): 640-649.

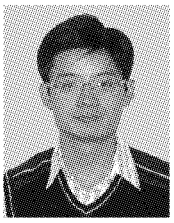
[7]TSANG E C C, CHEN D G, YEUNG D S, et al. Attribute reduction using fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2008, 16(5): 1130-1141.

[8]张志飞, 苗夺谦. 基于粗糙集的文本分类特征选择算法[J]. 智能系统学报, 2009, 4(5): 453-457.
ZHANG Zhifei, MIAO Duoqian. Feature selection for text categorization based on rough set[J]. CAAI Transactions on Intelligent Systems, 2009, 4(5): 453-457.

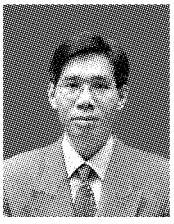
[9]SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems [M]//Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer Academic Publishers, 1992: 331-362.

[10]常犁云, 王国胤, 吴渝. 一种基于Rough Set理论的属性约简及规则提取方法[J]. 软件学报, 1999, 10(11): 1206-1211.
CHANG Liyun, WANG Guoyin, WU Yu. An approach for attribute reduction and rule generation based on rough set theory[J]. Journal of Software, 1999, 10(11): 1206-1211.

作者简介:



杨成东,男,1984年生,讲师,博士,主要研究方向为数据挖掘、粗糙集理论、智能计算.主持山东省高等学校科技计划项目等,发表学术论文十余篇.



邓廷权,男,1965年生,教授,博士生导师,主要研究方向为模糊信息分析、数学形态学与图像分析、智能识别与计算机视觉.主持国家自然科学基金、中国博士后科学基金、黑龙江省博士后科学基金等多项科研项目.近年来,发表学术论文30余篇,其中半数被SCI、EI、ISPT等检索.