DOI: 10.13656/j.cnki.gxkx.2009.04.013
广西科学 Guangxi Sciences 2009, 16(4): 385~ 388

决策系统属性约简的关系矩阵算法*

Attribute Reduction in Decision System Based on Relation Matrix Method

邓春燕[1,2],阮 忠[1],吕跃进[2]
DENG Chun-yan[1,2], RUAN Zhong[1], LV Yue-jin[2]

(1.河池学院计算机与信息科学系,广西宜州 546300; 2.广西大学数学与信息科学学院,广西南宁 530004)

(1. Department of Computer and Information Science, Hechi University, Yizhou, Guangxi, 546300, China; 2. Department of Mathematic and Information Science, Guangxi University, Nanning, Guangxi, 530004, China)

摘要:建立决策系统与关系矩阵之间的联系后,从关系矩阵的角度研究属性重要性的指标,并利用该指标作为启发式信息提出一种新的属性约简算法.该算法具有较大的灵活性,能够从搜索空间逐次删除不重要的属性,避免对这些属性进行重复搜索,提高了搜索的效率.该算法可行有效.

关键词:粗糙集 属性约简 相对约简 关系矩阵
中图法分类号:O159,TP301.6 文献标识码:A 文章编号:1005-9164(2009)04-0385-04

Abstract A novel algorithm for attribute reduction in decision system is proposed. For that, firstly our paper constructs the correlation between decision system and relation matrix. And then, studies formulas measuring attribute significance on the perspective of relation matrix, and uses the new formulas as heuristic information to design a new attribute reduction algorithm. Compared with the existing algorithms, the algorithm developed in this paper can avoid repeatedly search these attributes by gradually deleting unimportant attributes from searching space, thus the efficiency can be improved. Besides, a example was given to verify both the feasibility and efficiency of the algorithm.

Key words rough set, attribute reduction, relative reduction, relation matrix

粗糙集理论是一种新的处理模糊和不确定知识的数学工具.目前,它已被成功应用于人工智能、数据挖掘、模式识别与智能信息处理等领域,并越来越引起国际学术界的关注.属性约简是知识发现的重要课题,也是粗糙集理论中的核心内容之一.众所周知,数据集中的属性并不是同等重要的,甚至其中某些属性是冗余的,特别是随机采集数据的冗余度更大.冗余数据的存在,一方面是对资源(存储空间)的浪费;另一方面,干扰人们利用这些数据做出正确的决策.属性约简就是在保持数据分类或决策能力不变的条件下,删除其中冗余的或不重要的属性.

粗糙集理论中用决策表来表示含有决策属性的数据集,决策系统可能存在多个属性约简.因此,找到具有最少属性的约简(即最小约简)受到众多学者的关注和研究[1~5].然而,S. Wong和 W. Ziarko在1985年已经证明基于粗糙集的决策表最小约简是 NP-hard问题.导致属性约简是 NP-hard问题的主要原因是属性组合爆炸.因此利用数据集的特征,在算法中融入启发性知识,缩小问题求解的搜索空间则是一类被认为是更有意义的方法[2].

我们比较研究现有决策表的属性约简算法,以关系矩阵为基础,结合矩阵的行向量设计一个合理度量

收稿日期:2009-04-10
修回日期:2009-06-12
作者简介:邓春燕(1971-),女,硕士研究生,讲师,主要从事数据挖掘,粗糙集研究。
* 国家自然科学基金项目(70861001),广西研究生科研创新项目(2008105930701M51),广西河池学院科研项目(2008B-N003)资助。

属性重要性的公式,然后利用新公式作为启发式信息,提出一种基于关系矩阵的最小属性约简算法.

1 预备知识

粗糙集的基本概念参见文献 [6].

四元组 $S=(U,A,V,f)$ 是决策表,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 为论域; $A=C\cup D$, $C\cap D=\varnothing$, $C=\{c_1,c_2,\cdots,c_m\}$ 为条件属性集; $D=\{d_1,d_2,\cdots,d_k\}$ 为决策属性集, $V=\bigcup_{a\in A}V_a$, V_a 为属性 a 的值域; $f: U\times A\rightarrow V$ 是一个信息函数,即 $\forall a\in A, x\in U, f(x,a)\in V_a$. $\forall P\subseteq A$, P 决定了不可区分关系 $Ind(P)$, $Ind(P)=\{(x,y)\in U\times U|f(x,a)=f(y,a),\forall a\in P\}$, $Ind(P)$ 确定了 U 的一个划分,记为 U/P.

$X\subseteq U$, $P\subseteq C$, 则 X 的 P 下近似集 $\underline{P}X=\bigcup\{[x]_P|[x]_P\subseteq X\}$.

D 的 P 正域为 $Pos_P(D)=\bigcup\{Y\in U/P|Y\subseteq X, X\in U/D\}$.

$c_i\in C$,若 $Pos_C(D)=Pos_{C-\{c_i\}}(D)$,则称 c_i 为 C 中 D 不必要的,否则称 c_i 为 C 中 D 必要的.若 C 中的每一个 c_i 均为 D 必要的,则称 C 为 D 独立的.

$P\subseteq C$, P 是 D 的相对约简当且仅当 P 是 C 的 D 独立子族且 $Pos_P(D)=Pos_C(D)$.

C 中所有 D 必要的原始关系构成的集合称相对核,记为 $Core_D(C)$.

属性 $c_i\in C$ 关于属性集 D 的重要性为 $\sigma_{CD}(c_i)=\gamma_C(D)-\gamma_{C-\{c_i\}}(D)$,其中 $\gamma_C(D)$ 表示属性集 D 和 C 属性集之间的依赖度且 $\gamma_C(D)=|Pos_C(D)|/|U|$.

性质 1[7] 若 $\sigma_{CD}(c_i)>0$,则 $c_i\in Core_D(C)$.

定义 1 设 $f: U\times A\rightarrow V$ 是信息函数, $R^{(a)}=\{(x,y)\in U\times U|f(x,a)=f(y,a),a\in A\}$,称 $R^{(a)}$ 为 U 关于属性 a 的等价关系.其关系矩阵仍记为 $R^{(a)}=(r_{ij}^{(a)})_{n\times n}$.其中: $r_{ij}^{(a)}=\begin{cases}1,(x_i,x_j)\in R;\\0,(x_i,x_j)\notin R.\end{cases}$

性质 2[8] 设 $\{R_t|t\in T\}$ 是 X 上的一个 $FUZZY$ 关系集,若 $\forall t\in T$, R_t 是传递的,则 $\bigcap R_t(t\in T)$ 也是传递的.其中 $(A\cap B)_{(x)}=A(x)\wedge B(x)=\min(A(x),B(x))$.

若无特别说明,文中的符号含义为,(1) $R_i=(r_{i1},r_{i2},\cdots,r_{in})=(r_{ij})_{1\times n}$, $R_{ij}=r_{ij}$; (2) $Sum(R_i^{(C)}\leqslant R_i^{(D)})$ 表示关系矩阵 $R^{(C)}$ 的行向量小于或者等于关系矩阵 $R^{(D)}$ 对应行向量的总个数.

定义 2 在关系矩阵 $R=(r_{ij})_{n\times n}$ 中,对于 R_i, R_j,若 $\forall k\in\{1,2,\cdots,n\}$,有 $r_{ik}\leqslant r_{jk}$,则称 R_i 小于或等于 R_j.记为 $R_i\leqslant R_j$.

2 基于关系矩阵的决策系统属性约简算法

2.1 算法的理论依据

定义 3 设 $W=(w_{ij})_{n\times n}$, $T=(t_{ij})_{n\times n}$ 是 U 上两个不同的二元等价关系矩阵,令 $W\cap T=R=(r_{ij})_{n\times n}$,其中 $r_{ij}=\min(w_{ij},t_{ij})$.称 $W\cap T$ 为关系矩阵 W 与 T 的交.

由定义 1及性质 2易知,下列结论成立:

(1) 矩阵交运算满足交换律,即 $W\cap T=T\cap W$.

(2) 矩阵交运算满足结合律,即 $(W\cap T)\cap P=W\cap(T\cap P)$.

(3) $W\cap T$ 矩阵对应的二元关系仍为等价关系.

定义 4 设 C、D 分别为决策表的条件属性集和决策属性集. $\forall c_i\in C$, c_i 关于 D 的重要度定义为 $Sig_{CD}(c_i)=Sum(R_i^{(C)}\leqslant R_i^{(D)})-Sum(R_i^{(C-\{c_i\})}\leqslant R_i^{(D)})$.

定义 4说明 $Sig_{CD}(c_i)$ 的值越大,属性 c_i 在 D 中就越重要.我们利用 $Sig_{CD}(c_i)$ 作为启发式信息来寻找最小相对约简,以减少搜索空间.

引理 设 $R=(r_{ij})_{n\times n}$ 是 U 上一个二元等价关系的关系矩阵, $[x_i]_R$ 是 U 中由 x_i 生成的关于 R 的一个等价类,则 $\forall x_j\in[x_i]$ 当且仅当 $r_{ij}=1$.

由定义 1及等价类概念,易知引理成立.

定理 1 已知 $Pos_C(D)=\bigcup\{Y\in U/C|Y\subseteq X, X\in U/D\}$,记 $Pos_C'(D)=\{x_i\in U|R_{(i)}^{(C)}\leqslant R_{(i)}^{(D)}\}$,则有: $Pos_C'(D)=Pos_C(D)$. 即 $i\in\{1,2,\cdots,n\}$, $Pos_C'(D)$ 与 $Pos_C(D)$ 等价.

证明 (Ⅰ) 由正域定义 $Pos_C(D)=\bigcup\{Y\in U/C|Y\subseteq X, X\in U/D\}$ 得知: $\forall x_i\in Pos_C(D)$,必有 $[x_i]_C\subseteq[x_i]_D$.即 $\forall x_t\in[x_i]_C$,必有 $x_t\in[x_i]_D$.所以由引理得: $\forall x_t\in[x_i]_C$,有 $r_{it}^{(C)}=1$, $r_{it}^{(D)}=1$,即 $R_i^{(C)}\leqslant R_i^{(D)}$.因此 $Pos_C(D)\subseteq Pos_C'(D)$.

(Ⅱ) 设 $x_i\in Pos_C'(D)$,则由等价类性质知,至少存在一个 $j\in\{1,2,\cdots,n\}$,使得 $r_{ij}^{(C)}=1$.

因为 $R_i^{(C)}\leqslant R_i^{(D)}$,所以 $\forall r_{ij}^{(C)}=1$,必有 $r_{ij}^{(D)}=1$.亦即 $\forall x_i\in[x_j]_C$,必有 $x_i\in[x_j]_D$,因此 $x_i\in Pos_C(D)$.所以 $Pos_C'(D)\subseteq Pos_C(D)$.

综合(Ⅰ)、(Ⅱ)得: $Pos_C'(D)$ 与 $Pos_C(D)$ 是等价的.定理 1证明完毕.

定理 2 在决策系统 S 中, $c_i\in C$,若 $Sig_{CD}(c_i)>0$,则 $c_i\in Core_D(C)$.

证明 设 $Sum(R_i^{(C)}\leqslant R_i^{(D)})=n_0$, $Sum(R_i^{(C-\{c_i\})}\leqslant R_i^{(D)})=n_i$,

则由定理 1有: $|Pos_C(D)| = n_0$, $|Pos_{C-\{c_i\}}(D)| = n_i$.

又因为

$$Sig_{CD}(c_i) = Sum(R_i^{(C)} \leqslant R_i^{(D)}) - Sum(R_i^{(C-\{c_i\})} \leqslant R_i^{(D)}) = n_0 - n_i > 0,$$

即 $n_0 > n_i$.所以

$$\sigma_{CD}(c_i) = \gamma_C(D) - \gamma_C(D) = \frac{n_0}{|U|} - \frac{n_i}{|U|} > 0,$$

由性质 1知 $c_i \in Core_D(C)$.定理 2证明完毕.

定理 3 在决策系统 S中, $P \subseteq C$,如果 $\forall c_i \in P$,均有 $Sig_{PD}(c_i) > 0$,而且 $Sum(R_i^{(P)} \leqslant R_i^{(D)}) = n_0$,则 P是 C的 D约简,也是 C的 D最小约简.

证明 由定理 1得

$$POS_P(D) = \{x_i \in U | R_i^{(P)} \leqslant R_i^{(D)}, i \in \{1,2,\cdots,n\}\},$$

因为 $Sum(R_i^{(P)} \leqslant R_i^{(D)}) = n_0$, 即 $|Pos_C(D)| = |Pos_P(D)| = n_0$,所以 $Pos_P(D) = Pos_C(D)$. 又因为 $\forall c_i \in P$,有 $Sig_{PD}(c_i) > 0$,所以 c_i是必要的,因此 P是 D的一个相对约简.定理 3证明完毕.

2.2 算法的具体步骤

由定理 2可以求出 $Core_D(C)$,由定理 3可以判断 $Core_D(C)$是否是 D的相对约简.如果 $Core_D(C)$不能构成 D的相对约简,则可依次从非核的属性集 P中,选取 $Sig_{PD}(c_i)$值最大的属性 c_i,添加到 $Core_D(C)$中,直至能构成 D的相对约简为止.

基于关系矩阵的决策系统属性约简算法的步骤为,输入决策系统 S;输出 D的相对约简 P.

步骤 1 写出 U关于各属性的关系矩阵 $R^{(c_1)}$, $R^{(c_2)},\cdots,R^{(c_m)},R^{(d_1)},R^{(d_2)},\cdots,R^{(d_k)}$.

步骤 2 $R^{(C)} \leftarrow R^{(c_1)} \cap R^{(c_2)} \cap \cdots \cap R^{(c_m)}$, $R^{(D)} \leftarrow R^{(d_1)} \cap R^{(d_2)} \cap \cdots \cap R^{(d_k)}$, $T \leftarrow \varnothing$.

步骤 3 $\forall c_i \in C$,若 $Sig_{CD}(c_i) > 0$,则 $T \leftarrow T \cup \{c_i\}$.

步骤 4 $P \leftarrow T$,若 $P = C$,则转至步骤 7;否则转至步骤 5.

步骤 5 在 $(C - P)$中取 c_k, c_k满足 $Sig_{PD}(c_k) = \max\limits_{c_i \in C-P}(Sig_{PD}(c_i))$.

步骤 6 $P \leftarrow \cup \{c_k\}$, $C \leftarrow (C - \{c_k\})$.

步骤 7 若 $Sum(R_i^{(P)} \leqslant R_i^{(D)}) = sum(R_i^{(C)} \leqslant R_i^{(D)})$,转步骤 8;否则转步骤 5.

步骤 8 输出 P.

2.3 算法的时间复杂度分析

由 2.2算法可知,步骤 5不断删除不必要的属性,由定理 3知该操作是在不影响求属性约简的情况下,不断地缩小搜索空间,提高搜索效率,且所求出相对约简是决策系统属性的最小约简.最坏的情况是无核无冗余属性,而且各个属性的重要度 $Sig_{CD}(c_i)$均相等.此时需要计算 $|C|$个属性重要度.而计算 $Sig_{CD}(c_i)$的时间复杂度为 $O(|U|^2)$,因此整个算法的时间复杂度为 $O(|C||U|^2)$.这与文献 [2,5]算法的时间复杂度相同.

3 算例

以表 1的决策表 $S = (U, C \cup D, V, f)$为例.表 1中 $C = \{c_1, c_2, c_3\}$, $D = \{d\}$,求 D的一个最小约简.

表 1 决策表

Table 1 Dicision table

U	c_1	c_2	c_3	d
x_1	2	2	0	1
x_2	1	2	0	0
x_3	1	2	0	1
x_4	0	0	0	0
x_5	1	0	1	0
x_6	2	0	1	1

相关的关系矩阵如下:

$$R^{(c_1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$R^{(c_2)} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix},$$

$$R^{(c_3)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

$$R^{(D)} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix},$$

$$R^{(C)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$R^{(P_1)} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

$$R^{(P_2)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$R^{(P_3)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$R^{(C)} = R^{(c_1)} \cap R^{(c_2)} \cap R^{(c_3)}$, $R_1^{(C)} \leqslant R_1^{(D)}$, $R_4^{(C)} \leqslant R_4^{(D)}$, $R_5^{(C)} \leqslant R_5^{(D)}$, $R_6^{(C)} \leqslant R_6^{(D)}$,所以 $Sum(R_i^{(C)} \leqslant R_i^{(D)}) = 4 > |C|$.

令 $P_1 = C - \{c_1\}$, $P_2 = C - \{c_2\}$, $P_3 = C - \{c_3\}$,则有: $R^{(P_1)} = R^{(c_2)} \cap R^{(c_3)}$, $R^{(P_2)} = R^{(c_1)} \cap R^{(c_3)}$, $R^{(P_3)} = R^{(c_1)} \cap R^{(c_2)}$,由 $R^{(P_1)}$, $R^{(P_2)}$, $R^{(P_3)}$ 得 $Sum(R_i^{(P_1)} \leqslant R_i^{(D)}) = 1$, $Sum(R_i^{(P_2)} \leqslant R_i^{(D)}) = 4$, $Sum(R_i^{(P_3)} \leqslant R_i^{(D)}) = 4$, $Sig_{CD}(c_1) = Sum(R_i^{(C)} \leqslant R_i^{(D)}) - Sum(R^{(C-\{c_1\})} \leqslant R_i^{(D)}) = 4 - 1 = 3 > 0$,所以 $c_1 \in Core_D(C)$.

$R_1^{(P_2)} \leqslant R_1^{(D)}$, $R_4^{(P_2)} \leqslant R_4^{(D)}$, $R_5^{(P_2)} \leqslant R_5^{(D)}$, $R_6^{(P_2)} \leqslant R_6^{(D)}$,即 $Sum(R_i^{(P_2)} \leqslant R_i^{(D)}) = 4$,由定理3得: $P_2\{c_1, c_3\}$ 是 D的最小相对约简;同理可得 $P_3 = C - \{c_3\} = \{c_1, c_2\}$也是 D的最小相对约简.以上说明新提出的算法可行、有效.

4 结束语

我们在比较研究现有决策表的属性约简算法基础上,设计一个合理度量属性重要性的公式,然后利用该公式作为启发式信息并结合文献[2]提出的快速计算划分方法设计一种新的基于关系矩阵的决策系统属性约简算法. 新算法时间复杂度为 $O(|C|^2|U|)$. 由于该算法是基于关系矩阵的算法,因此具有编程简单、容易的特点.实例验证说明算法是可行、有效的.

参考文献:

[1] Miao Duoqian, Wang jue. An information-based algorithm for reduction of knowledge[J]. IEEE ICIPS 97, 1991: 1155-1158.

[2] 李金海,吕跃进.一种基于关系矩阵的信息系统属性约简算法[J].计算机工程与应用,2008,44(9): 147-149.

[3] 雷晓蔚.粗集理论的矩阵方法[J].计算机工程与应用,2006(17): 73-75.

[4] 苗夺谦,胡桂荣.知识约简的一种启发式算法[J].计算机研究与发展,1999,36(6): 681-684.

[5] 高学东,丁军.基于简化差别矩阵的属性约简算法[J].系统工程理论与实践,2006(6): 101-107.

[6] 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,2006.

[7] 刘清. Rough集及 Rough推理[M].北京:科学出版社,2005.

[8] 胡宝清.模糊数学理论基础[M].武汉:武汉大学出版社,2005.

(责任编辑:韦廷宗)

有害草本植物水葫芦可用于制取清洁燃料

世界十大有害草本植物(紫茎泽兰,薇甘菊,空心莲子草,豚草,毒麦,互花米草,飞机草,水葫芦,假高粱,布袋莲)之一的水葫芦(凤眼莲)在中国南方19个省市的广大水域随处可见。其生长繁殖十分迅速,在适宜条件下每天可增加20%~30%的生物量,8个月内就能从10棵增至60万棵,造成破坏生态、堵塞河道、影响航运、阻碍排灌、危害水电设施等严重的环境污染。水葫芦属于纤维素类生物质,它与稻草、秸杆等农业废弃物相比,具有三个突出特点:一是含水量高达95%左右;二是含有丰富的蛋白质;三是根部富集了许多重金属等污染物。最近中国科学家开展的一项研究发现,在微波联合碱作用下,水葫芦的有机成分:纤维素、半纤维素、木质素和蛋白质等都发生不同程度的降解,再经过纤维素酶作用后水解生成大量葡萄糖和木糖,这些糖经过进一步发酵可以用来制取燃料酒精、氢气、甲烷等清洁燃料,从而为水葫芦的能源化利用开辟出一条新途径。如何寻找高效廉价方法使水葫芦高效降解成可资利用的还原糖,是利用水葫芦发酵制取酒精的关键步骤和技术难点。

(据科学网)