文章编号：2095-4980(2016)04-0625-05

一种改进的无核信息系统属性约简算法

杨素敏[1,2]，蒙　洁[1]，张政保[2]，袁红丽[2]

(1.电子信息系统复杂电磁环境效应国家重点实验室，河南 洛阳 471003；2.军械工程学院 信息工程系，河北 石家庄 050003)

摘　要：针对无核信息系统的特点，基于互信息提出了一种新的启发式属性约简算法，该算法以增加属性后的互信息增量和属性自身的信息熵2项指标作为评价属性重要度的依据。实验结果表明，该算法避免了对于没有核属性的无核信息系统因随机选择初始属性造成计算复杂度增大的问题，并且属性约简效率提高，属性约简后的个数也相对较少。

关键词：粗糙集；属性约简；互信息；无核信息系统

中图分类号：TN911.2　　**文献标识码：**A　　**doi：**10.11805/TKYDA201604.0625

Attribute reduction algorithm for non-core information system

YANG Sumin[1,2], MENG Jie[1], ZHANG Zhengbao[2], YUAN Hongli[2]

(1.State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang Henan 471003, China; 2.Department of Information Engineering, Ordnance Engineering College, Shijiazhuang Hebei 050003, China)

Abstract: According to the characteristics of non-core information system, one new heuristic attribute reduction algorithm is proposed based on mutual information, in which the evaluation of attribute importance depends on two indexes, the increment of mutual information and the information entropy. When one attribute is added to the reduction sets, the attribute with the largest attribute importance is selected for the core attribute. This method can solve the problem of increased computational complexity caused by the randomly selected attributes. The simulation experiments indicate that the proposed algorithm is effective, which can not only improve the efficiency of attribute reduction, but also decrease the number of attribute reduction.

Key words: rough set; attribute reduction; mutual information; non-core information system

属性约简就是在保持信息系统分类能力不变的情况下，约去不必要的属性，是电子信息系统进行指标优选的关键环节。粗糙集理论(Rough Set Theory，RST)作为一种刻画不确定、不完整知识和数据的表达、学习、归纳的数学工具和方法，能够有效分析和处理不精确、不完整和不一致等各种信息，并从中挖掘出隐含的知识，揭示潜在的规律。与其他处理不确定和不精确问题理论的最显著区别是：它无需提供所需处理数据集合之外的任何先验信息，如统计学中的概率分布，模糊集理论中的隶属度等，能够客观描述或处理问题的不确定性。

然而对于一个信息系统来说，Wong S K M 和 Ziarko W[1]已经证明了属性约简和值约简都是 NP(Non-Deterministic Polynomial)问题，因此一般研究的是启发式约简算法，以获取最优或次优属性约简和值约简。文献[2–4]利用差别矩阵的方法进行属性约简算法的研究。研究人员发现，如果建立知识和信息之间的关系，从信息熵的角度考察属性约简，可以获得高效的属性约简算法。苗夺谦[5]等人提出的基于互信息的知识约简算法，是建立在条件属性对决策属性的互信息基础上的；颜艳[6]等人提出了一种基于互信息的粗糙集知识约简算法；贾平[7]等人提出了一种基于互信息增益率的属性约简算法，这些算法能够利用启发式信息减少搜索空间，尽可能缩短搜索时间，最终得到一个最优或近似最优解。文献[8–9]利用变精确度粗糙集模型，研究了不完备信息系统中的知识获取和约简问题。但目前基于互信息的属性约简算法，都是以决策表的相对核为起点，逐次选择最重要的属性添加到相对核中，直到满足核属性的互信息与条件属性集的互信息相等为止。

但在无核信息系统中，信息系统因缺少核属性，每选择一次属性后，都要重新计算一次互信息，计算的复杂度显著增大。为此本文依据无核信息系统的特点，提出了一种新的属性约简算法，该算法首先以互信息增量和属

收稿日期：2015-05-07；**修回日期：**2015-06-24
基金项目：国家实验室开放课题基金资助项目(2015K0304B)

性的信息熵建立属性重要度考核方法,并初选所有属性中属性重要度最大的属性作为核属性,每次增加新属性时,选择最大的互信息增量和属性重要度。该算法避免了随机初选属性增加时间复杂度和计算量过大的问题,且能够更快地得到属性约简结果,约简个数也相对较少,保证属性约简的效率和准确度,同时保证了约简属性的质量。实验仿真结果证明了所设计算法的有效性。

1 粗糙集基本概念

定义1:设 $S=(U,A,V,f)$ 为一信息系统。其中,$U=\left\{U_1,U_2,\cdots,U_{[U]}\right\}$ 为非空有限集,称为论域对象空间;$A=\left\{a_1,a_2,\cdots,a_n\right\}$ 为属性的非空有限集,称为属性集合;$V=\cup A_a$,其中 $a\in A$,V_a 为属性a的值域;$f:U\times A\to V_a$ 为信息函数,使得对于 $\forall x\in U$,当x取属性a时,其在 V_a 中具有唯一值。同时,对于 $\forall x\in U$,有序列 $C(c_1(x),c_2(x),\cdots,c_n(x))$ 和 $D(d_1(x),d_2(x),\cdots,d_n(x))$,$A=C\cup D,C\cap D=\phi$,则该信息系统称为决策表。其中 $c_1(x),c_2(x),\cdots,c_n(x)$ 称为条件属性集。

定义2:在给定的知识表示系 $S=(U,A,V,f)$ 中,对于任意属性集 $B\subseteq A$,不可分辨关系:

$$IND(B)=\left\{(x,y)\in U\times U:\forall a\in B(f(x,a)=f(y,a)\right\} \tag{1}$$

定义 3:在给定的知识表示系 $S=(U,A,V,f)$ 中,若 $P,Q\subseteq A$,则 Q 的 P 正域 $POS_p(Q)$ 定义为:

$POS_p(Q)=\underset{X\subseteq U/P}{\cup}\underline{R(X)}$,其中 $\underline{R(X)}$ 为 X 的下近似。

定义 4:设 U 是一个论域,P 和 $\underline{U}$ 为论域 U 上的 2 个等价关系族(即知识):

$U/ind(P)=\{X_1,X_2,\cdots,X_n\}$,$U/ind(Q)=\{Y_1,Y_2,\cdots,Y_m\}$,则 P,Q 在 U 上的子集的概率分布定义如下:

$$\begin{cases}\left[X:p\right]=\begin{bmatrix}x_1 & x_2 & \cdots & x_n\\ p(x_1) & p(x_2) & \cdots & p(x_n)\end{bmatrix}\\ \left[Y:p\right]=\begin{bmatrix}y_1 & y_2 & \cdots & y_m\\ p(y_1) & p(y_2) & \cdots & p(y_m)\end{bmatrix}\end{cases} \tag{2}$$

式中:$p(x_i)=\dfrac{\left|x_i\right|}{U},i=1,2,\cdots,n$;$p(y_i)=\dfrac{\left|y_j\right|}{U},j=1,2,\cdots,m$;符号 $\left|E\right|$ 表示 E 的基数。

定义 5:根据信息论,知识 P 的信息熵 $H(P)=-\sum_{i=1}^{n}p(x_i)\log p(x_i)$,知识 P 相对于 Q 的条件熵 $H(Q\left|P\right.)$ 为:

$$H(Q\left|P\right.)=-\sum_{i=1}^{n}p(x_i)\sum_{j=1}^{m}p(y_j\left|x_i\right.)\log p(y_j\left|x_i\right.) \tag{3}$$

知识 P 相对于 Q 的互信息 $I(P;Q)$ 为:

$$I(P;Q)=H(Q)-H(Q\left|P\right.) \tag{4}$$

定义 6:设 U 是一个论域,P 和 Q 为 U 上的 2 个等价关系族,若 $Ind(P)=Ind(Q)$,则 $H(P)=H(Q)$。

定义 7:等价关系族 P 独立(不依赖)的充分必要条件是:对任意 $R\in P$,都有 $H(R\left|P-\left\{R\right\}\right.)>0$。

定义 8:设 U 是一个论域,P 和 Q 为 U 上的 2 个等价关系族,$Q\subseteq P$ 是 P 的一个约简的充分必要条件是下列 2 个条件成立:a) $H(P)=H(Q)$;b) 对任意的 $q\in Q$,有 $H(q\left|P-\left\{q\right\}\right.)>0$。

2 基于互信息的属性重要度

粗糙集理论与信息熵的关系:熵度量了事件的不确定性,即信息源提供的平均信息量的大小;条件熵 $H(Q\left|P\right.)$ 度量了事件 P 发生的前提下,事件 Q 仍存在的不确定性;互信息 $I(P;Q)$ 代表了包含在事件 P 中关于事件 Q 的信息,即互信息度量了一个信源从另一个信源获取的信息量的大小。

在决策表中,关心的是哪些条件属性对于决策更重要,就要考虑条件属性和决策属性之间的互信息。文献[10]提出利用添加某个属性引起的互信息变化的大小反映该属性的重要程度。其定义如下:

$$SGF(a,R,Q)=I(Q;R\cup\{a\})-I(Q;R)=H(Q|R)-H(Q|R\cup\{a\}) \tag{5}$$

依据该公式选择的属性在值域中含有的值较多，从信息论的角度来讲，就是选择取值混乱的属性，而且对于决策属性所选取的属性也不一定是最重要的。

针对以上问题，文献[8]利用互信息增益率改进了属性重要度，其定义如下：

$$SGF(a,R,Q)=I(Q|R\cup\{a\})-I(Q|R)/H(a)=(H(Q|R)-H(Q|R\cup\{a\}))/H(a) \tag{6}$$

利用这种度量方法不仅考虑了在属性中添加属性之后互信息的增量，而且考虑了它自身的信息熵。当互信息增量相同时，$H(a)$越小，相应的属性重要度越大。但该算法计算过程中依赖于核属性，当没有核属性时，公式(6)变为：

$$SGF(a,R,D)=I(D)-I(D|a)/H(a)=I(a,D)/H(a) \tag{7}$$

为此对式(6)进行改进，首先利用式(7)计算每个属性的重要度，选取重要度最大的属性为核属性 R^*，此时重要度公式变为：

$$SGF(a,R^*,Q)=I(Q|R^*\cup\{a\})-I(Q|R^*)/H(Q|a)=(H(Q|R^*)-H(Q|R^*\cup\{a\}))/H(a) \tag{8}$$

3　一种改进的基于互信息的属性约简算法

依据式(8)，本文提出了一种新的属性约简算法，该算法首先初选所有属性中属性重要度最大的属性作为核属性，以互信息增量和属性的信息熵建立了属性重要度考核方法，每次增加新属性时，选择互信息增量和属性重要度都最大。算法的具体描述如下：

输入：一个相容的决策表系统 $S=(U,A,V,f)$，$A=C\cup D, C\cap D=\phi$，C 为条件属性集，D 是决策属性，U 为论域。

输出：决策表的一个约简 R。

计算条件属性集 C 与决策属性集 D 的互信息 $I(C;D)$；

利用式(7)计算所有条件属性的重要性，并将属性重要度最大的属性设为核属性 R^*；

令 $R=R^*$，对属性集 $R'=C-R, C'=C-R$进行如下操作：

1) 对每个属性 $c_i\in C'$，计算 $I(Q|R\cup\{a\})-I(Q|R)/H(a)$，从中选择数值最大的元素 a，如果有多个属性的属性重要度相同，则再比较它们的互信息值，选取互信息最大的属性加入属性约简集，则 $R=R\cup\{a\}$，$C'=C-R$。

2) 判断 $I(C;D)$ 与 $I(R;D)$，如果两者相同，则转到步骤 3)，否则步骤 1)。

3) R 即为决策表的一个相对约简，输出 R。

4　算法性能分析

该算法属性约简中每次增加都是属性重要度最大的，或者属性重要度相等时互信息变化量最大者，保证了每次添加的属性肯定是对决策效果最大的属性。如果原属性个数为 N，约简后的属性个数为 m，则经过 m 次循环后，就可以在完全保持信息系统分类能力不变的情况下必要属性选取完毕，算法的时间复杂度为 $O(m\times N)$。采用文献[8]中的算法计算时，由于随机选择一个属性作为核属性，如果所选属性不是最重要的属性，而且最重要的属性又最后被选择时，会出现所约简后的属性个数仍为 N，而且需要经过 N 次循环，在最坏的情况下，当核为空集时，每选择一次属性后就要重新计算一次互信息，故该算法的时间复杂度为 $O(N^2)$。可见所提出属性约简算法在保证信息系统分类能力不变的情况下，属性约简率提高了 $1-m/N$。

5　仿真实验

为了验证算法的有效性，本文以无核信息系统样本为例，如表 1 所示。

该无核信息系统有 7 个条件属性，为 $C=\{c_1,c_2,c_3,c_4,c_5,c_6,c_7\}$，10 个样本 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}$，决策属性为 D。评估指标集 C 的值域为 $V=\{1,2,3,4\}$。决策属性集 D 的值域为{1,2,3}。按照第 3 节所描述的算法，具体约简步骤如下：

1) 按照式(1)，计算得到的决策属性集不可分辨集为：

$$IND(D)=\{\{x_1,x_2,x_4,x_7,x_9\},\{x_3,x_6,x_8,x_{10}\},\{x_5\}\}$$

2) 依据式(4)计算条件属性集 C 相对于决策集 D 的互信息：

$$I(C;D)=H(D)-H(D|C)=3.3219$$

3) 依据式(7)计算条件属性集 C 中每个属性的属性重要度，结果如表 2 所示，从表中选取属性重要度最大的属性 c_3 为核属性，即 $R^*=\{c_3\}$。

表 1 无核信息系统决策表

Table1 Decision table for non-core information system

U	c_1	c_2	c_3	c_4	c_5	c_6	c_7	D
x_1	2	2	3	1	1	4	1	1
x_2	1	1	4	2	2	2	2	1
x_3	2	4	2	2	2	4	2	2
x_4	2	1	4	3	2	2	3	1
x_5	4	4	1	3	2	4	3	3
x_6	2	2	3	3	3	3	3	2
x_7	1	1	4	3	2	1	3	1
x_8	2	2	2	2	2	2	4	2
x_9	3	3	2	4	4	3	4	1
x_{10}	3	3	2	2	2	2	4	2

表 2 属性互信息和属性重要度数值表

Table2 Mutual information and importance for attribute

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
mutual information	1.075 488 75	1.295 461 844	1.321 928 095	0.797 416 845	0.342 609 804	0.770 950 594	0.770 950 594
importance	0.610 7	0.657 3	0.715 9	0.463 1	0.252 5	0.417 5	0.417 5

4) 令 $R=R^*$，对属性集 $R^{'}=C-R, C^{'}=C-R$，进行如下操作：

对剩余的每个属性 $c_i\in C^{'}$，计算 $\{c_i,c_3\}^{'}$ 的属性重要度，结果如表 3 所示，从表中可以看出 c_4,c_5,c_6 3 个属性的属性重要度相同，而 c_6 的互信息最大，因此选取属性 c_6 加入属性约简集，则 $R=R\cup\{c_6\}$，$C^{'}=C-R$。

表 3 属性互信息和属性重要度数值表

Table3 Mutual information and importance for attribute

	$\{c_1,c_3\}$	$\{c_2,c_3\}$	$\{c_4,c_3\}$	$\{c_5,c_3\}$	$\{c_6,c_3\}$	$\{c_7,c_3\}$
mutual information	2.321 9	1.914 2	2.646 4	1.706 6	2.921 9	2.287 9
importance	0.920 7	0.905 4	1.000 0	1.000 0	1.000 0	0.864 5

5) $I(R;D)=2.9219$，此时 $I(R;D)\neq I(C;D)$；

6) 按照第 3 节的算法，需要从剩余属性中选择 1 个加入，对剩余的每个属性 $c_i\in C^{'}$，计算 $\{c_i,c_3,c_6\}$ 的属性重要度，结果如表 4 所示。

表 4 属性互信息和属性重要度数值表

Table4 Mutual information and importance for attribute

	$\{c_1,c_3,c_6\}$	$\{c_2,c_3,c_6\}$	$\{c_4,c_3,c_6\}$	$\{c_5,c_3,c_6\}$	$\{c_7,c_3,c_6\}$
mutual information	3.321 9	3.021 9	3.121 9	2.921 9	3.121 9
importance	1.000	0.968	1.000	1.000	1.000

按照第 3 节算法规则，选取属性 c_1 加入属性约简集，则 $R=R\cup\{c_6,c_1\}$，$C^{'}=C-R$。

7) 计算，此时 $I(R;D)=I(C;D)$，算法终止。$R=\{c_1,c_3,c_6\}$ 是原无核信息系统决策表的一个约简。

按照文献[8]所提的属性约简算法，由于没有考虑无核属性的因素，得到的属性约简为 $\{c_1,c_2,c_3,c_4,c_5,c_6,c_7\}$，可见采用本算法后，在保持信息系统分类能力不变的情况下，属性约简的数量减少了 57%，属性约简的时间大概为原有的 42%。通过本实例的实验结果可以看出，本文所提的算法对于无核系统有很好的属性约简质量和效率。

6 结论

本文针对无核信息系统缺少核属性的特点，提出了一种新的启发式属性约简算法，该算法利用互信息增量和属性的信息熵的比，考核属性的重要度。首先把所有属性中属性重要度最大的属性假定为核属性，每次增加新属性时，选择互信息增量和属性重要度都最大，保证了每次添加的属性肯定是对系统决策影响最大的属性。该算法避免了随机初选属性增加时间复杂度、计算量过大以及属性约简数量多的问题，不仅保证了属性约简结果的精简，而且也提高了属性约简的效率和准确度，确保了约简属性的质量。仿真实验验证了所设计算法的正确性和有效性。

参考文献:

[1] WONG S K M,ZIARKO W. On optional decision rules in decision tables[J]. Bulletin of Polish Academy of Sciences, 1985, 33(11/22):693–696.

[2] SKOWRON A,RAUSZER C. The discernibility matrices and functions in information systems[J]. Theory & Decision Library, 1992,11:331–362.

[3] YANG Ming,SUN Zhihui. Improvement of discernibility matrix and the computation of a core[J]. Journal of Fudan University (Natural Science), 2004,43(5):865–868.

[4] CHENG Jing,ZHU Jing,ZHANG Fan. An updated algorithm for attribute reduction based on discernibility matrix[J]. Journal of Hunan University(Natural Sciences), 2009,36(4):86–88.

[5] 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. (MIAO Duoqian,HU Guirong. A heuristic algorithm for reduction of knowledge[J]. Journal of Computer Research and Development, 1999,36(6):681–684.)

[6] 颜艳,杨慧中. 一种基于互信息的粗糙集知识约简算法[J]. 清华大学学报(自然科学版), 2007,47(S2):1903–1906. (YAN Yan,YANG Huizhong. Knowledge reduction algorithm based on mutual information[J]. Journal of Tsinghua University (Science and Technology), 2007,47(S2):1903–1906.)

[7] 贾平,代建华,潘云鹤,等. 一种基于互信息增益率的新属性约简算法[J]. 浙江大学学报(工学版), 2006,40(6):1041–1044. (JIA Ping,DAI Jianhua,PAN Yunhe,et al. Novel algorithm for attribute reduction based on mutual-information gain ratio[J]. Journal of Zhejiang University(Engineering Science), 2006,40(6):1041–1044.)

[8] 张明. 粗糙集理论中的知识获取与约简方法的研究[D]. 南京:南京理工大学, 2012. (ZHANG Ming. Research on knowledge acquisition and reduction in rough set theory[D]. Nanjing,China:Nanjing University of Science & Technology, 2012.)

[9] MA Minghua,DENG Tingquan. The attribute reduction of the information system based on new rough set[C]// 2011 2nd International Conference on ICICIP. Harbin,China:IEEE, 2011:301–304.

[10] 谭宗凤,徐章艳,王帅. 一种改进的粗糙集权重确定方法[J]. 计算机工程与应用, 2012,48(18):115–118. (TAN Zongfeng, XU Zhangyan,WANG Shuai. Improved method of attribute weight based on rough sets theory[J]. Computer Engineering and Applications, 2012,48(18):115–118.)

作者简介:

杨素敏(1971–),女,河北省藁城市人,博士,副教授,研究方向为计算机应用.email: yangaumin1971@sina.com.

蒙　洁(1972–),女,河南省洛阳市人,高工,研究方向为电子信息系统效能评估理论.

张政保(1965–),男,石家庄市人,教授,研究方向为计算机应用.

袁红丽(1982–),女,河北省栾城市人,讲师,研究方向为计算机应用.