一种新的决策粗糙集启发式属性约简算法

常红岩 蒙祖强
(广西大学计算机与电子信息学院 南宁 530004)

摘 要 属性约简是粗糙集理论中最重要的研究内容之一。在决策粗糙集中,学者提出了多种属性约简的定义,其中包括保持所有对象正决策不变的约简定义。针对该约简定义,为了高效地获取约简集,设计了一种启发式函数——决策重要度,这种启发式函数根据每个属性正决策对象集合的大小来定义其重要性,正决策对象集合越大表示重要性越高,由此构造了基于决策重要度的启发式属性约简算法。该算法的优点是通过对属性决策重要度的排序,确定了一个搜索方向,避免了属性的组合计算,减少了计算量,能够找出一个较小的约简集。实验结果表明,该算法是有效的,能够得到较好的约简效果。

关键词 决策粗糙集,属性约简,启发函数
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.044

New Heuristic Algorithm for Attribute Reduction in Decision-theoretic Rough Set

CHANG Hong-yan MENG Zu-qiang
(College of Computer, Electronics and Information, Guangxi University, Nanning 530004, China)

Abstract Attribute reduction is one of the most important research contents in rough set theory. Scholars have proposed various definitions for attribute reduction in decision-theoretic rough set, including the definition of keep the positive decisions of all objects unchanged. Directing at the positive decision definition, in order to efficiently obtain the reduction set, designed a heuristic function is designed, that is important degree of decision-making. This heuristic function defines the decision important degree of every attribute according to the size of positive decision objects set. The bigger the size of positive decision objects set, the greater the improtance, thus constructs heuristic attribute reduction algorithm based on the decision important degree. The advantage of this algorithm is that it determines the search direction according to the sorting of attribute decision important degree, avoids the calculation of attribute combination, and can reduce the amount of calculation and find out a smaller reduction set. The experimental results show that the algorithm is effective and can obtain a good reduction effect.

Keywords Decision-theoretic rough set, Attribute reduction, Heuristic function

1 引言

粗糙集理论(Rough set theory, RS)是1982年由波兰数学家Pawlak提出的,是一种处理不精确(imprecise)、不一致(inconsistent)、不完整(incomplete)等各种不完备信息的有效的数据分析理论,其主要优点之一就是不需要先验知识。目前经典Pawlak粗糙集在智能领域的机器学习、数据挖掘、决策支持和模式识别与分类等方面得到了广泛的应用,推动了人工智能的迅速发展。然而,由于经典粗糙集的上近似集和下近似集严格的包含关系和现实数据的不完备不一致性,导致Pawlak粗糙集缺乏容错能力。为了扩展粗糙集模型,Yao等学者在粗糙集中引入了贝叶斯理论并于20世纪90年代初提出了决策粗糙集理论(Decision-theoretic rough set, DTRS)[1]。它通过阈值增加了模型的容错能力,并且讨论了根据贝叶斯风险进行决策的过程。

基于启发式的约简在经典粗糙集中有较多的研究。苗等在文献[2]、沈玮等人在文献[3]以及施化吉等人在文献[4]中都从信息的角度,对决策表的属性重要性给出度量,在此基础上提出基于互信息的启发式约简算法。但该启发式算法对于求取相对约简是不完备的,即最后求得的相对约简中仍有冗余属性存在的可能。文献[5]提出了把可辨识矩阵和互信息相结合的启发式约简算法,先由可辨识矩阵计算出属性的核,以最大化相对互信息为原则,向核属性中增加新属性,直到得到一个约简集。但是,决策粗糙集不存在属性核,所以这种启发式的算法具有局限性,只适合用在经典粗糙集中,而不适合用在决策粗糙集中而且使用可辨识矩阵浪费存储空间。张燕在文献[6]中提出了基于边界域率的启发式属性约简算法,但是该算法只适用于相容决策表,对于不相容决策表需要经过

到稿日期:2015-05-14 返修日期:2015-08-24 本文受国家自然科学基金项目(61363027),广西自然科学基金(2012GXNSFAA053225)资助。
常红岩(1989—),女,硕士生,主要研究方向为数据挖掘、粗糙集,E-mail:1432108145@qq.com;蒙祖强(1974—),男,博士,教授,博士生导师,主要研究方向为人工智能、粒度计算、知识发现等。

预处理才可以求约简集，而且该算法的约简率较低，并且对于样本数量大的样本约简效果明显，而对于样本数量小的样本来说约简效果不明显。文献[7-9]是保持正域非减的约简，其中文献[7]提出以正域对所有样本数量的比值为属性重要度的启发式约简算法，由于在决策粗糙集中，正域的单调性已不再保持，因此基于正域的约简在可解释性上也存在一定的问题。

综上所述，由于在决策粗糙集模型中的上下近似集的定义中引入了概率包含关系，其正域的单调性特征不再保持，导致属性约简的定义和性质较经典粗糙集中属性约简有较大差异。已有的启发式函数基本都是根据正域的大小作为控制约简的条件，由于正域的单调性不保持，因此导致求得的约简仍有冗余属性存在的可能。文献[10]对几种属性约简之间的定义进行了研究，指出了保持所有对象的正决策不变的约简呈现稳定性，但是并没有给出具体的算法和实验分析证明其约简效果。因此本文针对正决策保持的约简定义，探讨了其单调性，为了减少属性的搜索空间，提出新的启发式函数，并提出一个基于启发函数的正决策保持的属性约简算法。

2 决策粗糙集理论相关的基本概念

本节主要定义粗糙集理论中的一些基本概念，详细介绍请参考文献[1,13]。

一个信息表知识表达系统 S 可以表示为：$S=(U,R,V,f)$。其中，U 是对象的集合，也称为论域，$R=C\cup D$ 是属性集合，子集 C 和 D 分别为条件属性集合结果属性集，$V=\bigcup_{a\in A}V_a$ 表示属性值的集合，V_a 是属性 a 的值域；$f:U\times R\to V$ 是一个信息函数，表示 U 中每个对象 x 的属性值。在不引起混淆的情况下，$S=(U,R,V,f)$ 一般简记为 (U,R)。

对于每个属性子集 $B\subseteq R$，定义一个不可分辨二元关系 $IND(B)$，即 $IND(B)=\{(x,y)\mid(x,y)\in U\times U,\ \forall b\in B,\ f(x,b)=f(y,b)\}$。显然，$IND(B)$ 是一个等价关系，且构成 U 的一个划分，简记为 U/B 或 π_B；令 $[x]_B=\{y\mid\forall b\in B,\ f(x,b)=f(y,b)\}$，$[x]_B$ 是 U/B 以 x 为代表元的等价类；条件属性集 C 导出的 U 上的划分记为 $\pi_C=\{C_1,C_2,\cdots,C_n\}$，决策属性 D 导出的 U 上的划分 $\pi_D=\{D_1,D_2,\cdots,D_m\}$。通过等价关系，可以描述 U 上的任何一个子集。

定义 1[1] 设子集 $X\subseteq U$，那么 X 关于属性集 B 的上近似和下近似定义为：

$$\overline{apr_B}(X)=\{x\mid x\in U\wedge[x]\cap X\neq\varnothing\}$$

$$\underline{apr_B}(X)=\{x\mid x\in U\wedge[x]\subseteq X\}$$

基于上近似和下近似，下面引入正域 $POS_B(X)$、负域 $NEG_B(X)$ 和边界域 $BND_B(X)$ 的定义：

$$POS_B(X)=\underline{apr_B}(X)$$

$$NEG_B(X)=U-\overline{apr_B}(X)$$

$$BND_B(X)=\overline{apr_B}(X)-\underline{apr_B}(X)$$

接下来引入决策粗糙集中的相关概念，详细介绍见参考文献[1]。

当对象 x 属于 X 时，令 λ_{PP}，λ_{NP} 和 λ_{BP} 代表将一个对象相应划分到集合 $POS_\pi(X)$，$NEG_\pi(X)$ 和 $BND_\pi(X)$ 中时的损失函数。反之，当 x 不属于 X 时，λ_{PN}，λ_{NN} 和 λ_{BN} 代表采取相应动作时的损失函数。通常损失函数满足 $\lambda_{PP}\leqslant\lambda_{BP}\leqslant\lambda_{NP}$ 和 $\lambda_{NN}\leqslant\lambda_{BN}\leqslant\lambda_{PN}$。

$$\alpha=\frac{(\lambda_{PN}-\lambda_{BN})}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}$$

$$\gamma=\frac{(\lambda_{PN}-\lambda_{NN})}{(\lambda_{NP}-\lambda_{PP})+(\lambda_{PN}-\lambda_{NN})}$$

$$\beta=\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}$$

在 $0<\beta<\alpha<1$ 时，贝叶斯决策过程可以表示为：

(P) 如果 $P(X|[A])\geqslant\alpha$，则 $x\in POS_\pi(X)$

(N) 如果 $P(X|[A])\leqslant\beta$，则 $x\in NEG_\pi(X)$

(B) 如果 $\beta<P(X|[A])<\alpha$，则 $x\in BND_\pi(X)$

一个等价类 $[x]_A\in\pi_A$ 属于决策类 D_i 的条件概率定义为：

$$p(D_i|[x]_A)=\frac{|[x]_A\cap D_i|}{|[x]_A|}$$

那么，基于条件属性集 A 的划分在 π_D 下的正区域、边界域和负区域定义为：

$$POS_{(\alpha,\beta)}(\pi_D|\pi_A)=\{x\in U|P(D_{\max}([x]_A)|[x]_A)\geqslant\alpha\}$$

$$BND_{(\alpha,\beta)}(\pi_D|\pi_A)=\{x\in U|\beta<P(D_{\max}([x]_A)|[x]_A)<\alpha\}$$

$$NEG_{(\alpha,\beta)}(\pi_D|\pi_A)=\{x\in U|P(D_{\max}([x]_A)|[x]_A)\leqslant\beta\}$$

其中，$D_{\max}([x]_A)=\underset{D_i\in\pi_D}{\operatorname{argmax}}\left\{\frac{|[x]_A\cap D_i|}{|[x]_A|}\right\}$，表示等价类被划分到具有最大概率的决策类。

可以看出，在决策粗糙集模型中，论域 U 可以通过阈值 α 和 β 划分为 3 个区域，即总体决策类正域、总体决策类负域、总体决策类边界域。从语义上看，这 3 个区域可以对应 3 种规则类型及三支决策语义，即正域决策、负域决策、边界域决策，具体描述如下。

若在 $[x]_A$ 的描述下，X 发生的概率大于阈值 α，则将 $[x]_A$ 划分到 X 的正域中，执行正域决策。

若在 $[x]_A$ 的描述下，X 发生的概率介于阈值 α 和 β 之间，则将 $[x]_A$ 划分到 X 的边界域中，此时决策依据不足，暂不决策，需收集更多信息以便做出正确决策。

若在 $[x]_A$ 的描述下，X 发生的概率小于阈值 β，则将 $[x]_A$ 划分到 X 的负域中，执行负域决策。

3 一种新的启发式函数

3.1 决策粗糙集中的属性约简

在 Pawlak 粗糙集模型中，正区域大小随着属性增多而增大，具有单调性特征；而决策粗糙集中，α 正域已经不再具备严格的单调性特征。由于属性约简是决策粗糙集最重要的研究内容之一，目前学者提出了保持概率正区域不变的约简[7]、保持分类质量不变的约简[10]和保持局部最大概率正区域的约简[11]等多种属性约简模型。

但是保持概率正区域不变的约简和保持分类质量不变的约简随着属性个数的减少经常出现跳跃性，使约简具有不稳定性，即约去某些属性后是决策表的约简，再约去一个属性后不是原决策表的约简，但是继续约简又可以得到原决策表的约简。

Yao 等人在文献[1]中提出了正域约简的定义，其具体描述如定义 2 所示。

定义 2 设 $S=\{U,C\cup D,f,V\}$ 为一个决策表，$\alpha\in[0,1]$，若属性子集 $B\subseteq C$ 满足以下两个条件：

(1)正域非减性，$|POS_B^{\alpha}(D)|\geqslant|POS_C^{\alpha}(D)|$；

(2)属性独立性，对任意 $a\in B$，$|POS_{B-\{a\}}^{\alpha}(D)|<|POS_B^{\alpha}(D)|$。

则称属性子集 B 为属性全集 C 的一个决策粗糙集 α-正域约简。

文献[10]指出，当概率正区域中对象增多时，其代价也越大。从实际应用来说，由于放宽了条件，将原来不符合条件的对象变成符合条件的对象，必定付出一定的代价。因此，我们需要提出新的约简概念。

3.2 基于正决策的约简

Zhao 等人[10]提出了保持所有对象正决策不变的约简，该约简方法能够保证一个属性如果不可以约简，那么在整个约简过程中都是不可约简的，所以使约简具有稳定性，使约简随着属性个数的减少不会出现跳跃现象。本文就是根据保持所有对象正决策不变的理论，提出新的启发式函数，继而提出新的启发式属性约简算法。

对于任意一个等价类 $[x]_A\in\pi_A$，令 $D_{POS(\alpha,\beta)}$ 代表所有对象的正决策对象的集合，是当阈值大于或等于 α 时，对应的决策类的集合。则其描述如下[10]：

$$D_{POS(\alpha,\beta)}([x]_A)=\{D_i\in\pi_D\mid P(D_i\mid[x]_A)\geqslant\alpha\}$$

定义 3[10] 给定决策表 S，属性子集 $A\subseteq C$，若 A 满足以下两个条件：

(1) $\forall x\in U, D_{POS(\alpha,\beta)}([x]_A)=D_{POS(\alpha,\beta)}([x]_C)$

(2) $\forall a\in U, D_{POS(\alpha,\beta)}([x]_{A-\{a\}})\neq D_{POS(\alpha,\beta)}([x]_A)$

则称属性子集 B 为属性全集 C 的一个正决策约简(记为 Red_{Dpos})。

其中，$D_{POS(\alpha,\beta)}([x]_C)$ 代表属性全集 C 所决定的正决策对象的集合，作为 U 的一个约简集 A，应该与属性全集具有相同的正决策对象的集合。定义 3 中的条件(1)能够保证 A 是一个约简，但不能保证 A 是一个最小的约简。条件(2)检查 A 中的所有属性，删除了冗余属性，能够保证得到一个最小的约简。

正决策保持的约简的特点是比较的是决策类中的对象集合。由于同一决策类中对象类别都是一样的，能够保证快速找到和属性全集具有同样分类能力的约简集。而正区域非减的约简，比较的是正区域中的对象，由于不一致对象的存在，会导致最终约简集的分类正确率有所下降。

定理 1[10] 给定决策表 S，$\forall x\in U, a\in C$，若 $D_{POS(\alpha,\beta)}([x]_{C-\{a\}})\neq D_{POS(\alpha,\beta)}([x]_C)$，则 $\forall B\subseteq C-\{a\}$，均有 $D_{POS(\alpha,\beta)}([x]_B)\neq D_{POS(\alpha,\beta)}([x]_C)$。

定理 1 表明，保持所有对象的正决策不变的属性约简方法能够保证一个属性若不可约简，则在整个约简过程中都是不可约简的，从而保证约简具有稳定性。

3.3 启发函数的定义

启发式约简算法一般是先根据启发函数确定一个搜索方向，以此来达到减少计算量并找出一个约简集的目的。针对决策粗糙集，正域不再随着属性的增加而增大，因此不适合以正域为终止往约简集中增加属性的条件，所以需要提出新的启发搜索的终止条件，即保持正决策不变。同时，也有必要提出新的启发式函数作为判断属性重要性的依据。

由 $D_{POS(\alpha,\beta)}([x]_a)=\{D_i\in\pi_D\mid P(D_i\mid[x]_a)\geqslant\alpha\}$ 可知，如果 a 在约简集中，那么以 a 对样本进行等价类划分得到的 $[x]_a$ 中各个对象的决策属性取值相等的可能性较大，也就是说 $[x]_a$ 与某个决策类 D_i 的交集较大，也就意味着属性 a 对于决策的影响比较大。这说明了属性 a 的正决策对象集合的大小可以作为 a 对决策影响的大小，成为属性重要度的判定方法，用于构造启发式函数。

定义 4 给定决策表 S，$\alpha\in(0.5,1]$ 为条件概率阈值，$a\in C$ 为单个属性，则属性 a 的 α-决策重要度定义为 $\gamma_a^{\alpha}=\dfrac{|D_{POS_{(\alpha,\beta)}}([x]_{\{a\}})|}{|U|}$。其意义是条件属性 a 对于决策属性的所有正决策对象集合的大小在整个论域中的比例，比例越大，说明属性 a 的正决策对象的个数越多，属性 a 的决策重要度也越大，同时也说明了 a 的分类能力强。

本文以此决策重要度函数作为启发式函数，应用于构造属性约简算法。用此决策重要度作为启发函数的优点是决策重要度在约简算法中有较小的时间复杂度，计算所有属性的决策重要度的时间复杂度为 $O(|C||U|^2)$，而在其他的约简算法中计算启发信息的时间复杂度基本上会达到 $O(|C|^2|U|^2)$。

4 一种新的决策粗糙集约简算法

下面给出以属性的决策重要度为启发函数的保持所有对象正决策不变的启发式属性约简算法。该算法首先计算所有条件属性的 α-决策重要度，然后按照决策重要度从大到小排序；依次将决策重要度较大的条件属性加入到约简集 R 中，直到集合 R 满足定义 3 中的条件(1)为止；逐一检查 R 中的每一个条件属性 c 是否满足条件 $D_{POS(\alpha,\beta)}([x]_{R-\{c\}})=D_{POS(\alpha,\beta)}([x]_R)$，若满足，则从 R 中删去 c。其中计算等价类本文使用的是文献[12]的高效算法，添加属性和删除冗余属性使用的是文献[14]中的算法框架。

以决策重要度为启发函数的保持所有对象正决策不变的启发式属性约简算法如算法 1 所示。

算法 1

输入：一个决策表 S，概率包含度阈值 α

输出：该决策表的一个约简 R

(1)计算 C 中每个属性 a 的决策重要度 $\gamma_a^{\alpha}=\dfrac{|D_{POS_{(\alpha,\beta)}}([x]_{\{a\}})|}{|U|}$；

(2)将属性按照决策重要度由大到小排列，令为 Rank；

(3)令初始约简集 R 为 Rank 中的第一个属性 a，将 a 从 Rank 中删除，Rank=Rank−{a}；

(4)在 $D_{POS(\alpha,\beta)}([x]_R)\neq D_{POS(\alpha,\beta)}([x]_C)$ 时，重复如下循环：

令 a 为 Rank 中第一个属性，将 a 并入 R，R=R∪{a}；同时将 a 从 Rank 中删除；

(5)对所有的 c∈R，进行如下循环：若 $D_{POS(\alpha,\beta)}([x]_{R-\{c\}})=D_{POS(\alpha,\beta)}([x]_R)$，则 R=R−{c}；

(6)输出该决策表的一个保持所有对象正决策不变的约简 R。

第(1)步中计算所有属性的决策重要度的时间复杂度是 $O(|C||U|^2)$，其中 $|C|$ 表示属性的个数，$|U|$ 表示论域的个数；在第(4)步中，往约简集 R 中添加属性，需要计算约简集 R 的正决策对象集合，时间复杂度为 $O(|R||C||U|^2)$，其中 $|R|$ 代表约简集中属性的个数；第(5)步和第(4)步的时间复杂度是一样的。所以，总的来说，该算法的时间复杂度是 $O(|R||C||U|^2)$。

为了验证算法 1 的有效性，将算法 1 与文献[7]中的关于保持正域非减的约简算法结果进行了比较。实验中令 $\alpha=0.6$。

考虑决策表 $S=(U,C\cup D)$，其中，$U=\{x_1,x_2,\cdots,x_9\}$，$C=\{a_1,a_2,\cdots,a_6\}$，$D=\{d\}$，如表 1 所列。

表 1　决策表

U	a_1	a_2	a_3	a_4	a_5	a_6	d
x_1	1	1	1	1	1	1	1
x_2	1	0	0	0	1	1	1
x_3	0	1	1	1	0	0	2
x_4	1	1	1	0	0	1	2
x_5	0	0	1	1	0	1	2
x_6	1	0	0	0	1	1	3
x_7	0	0	1	1	1	0	3
x_8	1	0	0	0	1	1	3
x_9	0	0	1	1	0	1	3

计算论域 U 在属性集 $\{d\}$ 下的划分为 $U/\{d\}=\{\{x_1,x_2\},\{x_3,x_4,x_5\},\{x_6,x_7,x_8,x_9\}\}$，论域 U 在属性集 C 上的划分为 $U/\{C\}=\{\{x_1\},\{x_2,x_6,x_8\},\{x_3\},\{x_4\},\{x_5,x_9\},\{x_7\}\}$，则可求出在属性集 C 上的 $D_{POS(\alpha,\beta)}([x]_C)=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9\}$。

下面分别计算 U 对属性 a_1,a_2,a_3,a_4,a_5,a_6 的划分分别为：

$U/\{a_1\}=\{\{x_1,x_2,x_4,x_6,x_8\},\{x_3,x_5,x_7,x_9\}\}$

$U/\{a_2\}=\{\{x_1,x_3,x_4\},\{x_2,x_5,x_6,x_7,x_8,x_9\}\}$

$U/\{a_3\}=\{\{x_1,x_3,x_4,x_5,x_7,x_9\},\{x_2,x_6,x_8\}\}$

$U/\{a_4\}=\{\{x_1,x_3,x_5,x_7,x_9\},\{x_2,x_4,x_6,x_8\}\}$

$U/\{a_5\}=\{\{x_1,x_2,x_6,x_7,x_8\},\{x_3,x_4,x_5,x_9\}\}$

$U/\{a_6\}=\{\{x_1,x_2,x_4,x_5,x_6,x_8,x_9\},\{x_3,x_7\}\}$

则可求出每个属性的正决策对象的集合分别为：

$D_{POS(\alpha,\beta)}([x]_{a1})=\emptyset$

$D_{POS(\alpha,\beta)}([x]_{a2})=\{x_3,x_4,x_5,x_6,x_7,x_8,x_9\}$

$D_{POS(\alpha,\beta)}([x]_{a3})=\{x_6,x_7,x_8,x_9\}$

$D_{POS(\alpha,\beta)}([x]_{a4})=\emptyset$

$D_{POS(\alpha,\beta)}([x]_{a5})=\{x_3,x_4,x_5,x_6,x_7,x_8,x_9\}$

$D_{POS(\alpha,\beta)}([x]_{a6})=\emptyset$

根据决策重要度排序的结果是：a_2,a_5,a_3,a_1,a_4,a_6。

根据算法 1 得到的决策表 1 约简集为 $\{a_2,a_5\}$。

然而，根据文献[7]中的保持正域非减的算法得到的决策表 1 的约简为 $\{a_2\}$。由此可以看出，本文算法得到的约简集明显优于文献[7]得到的约简集，因为直观上就可以看出以单个属性 a_2 作为约简集是不合理的。

根据得到的约简分别重新对决策表 1 的 9 条记录进行分类，用本文算法 1 的得到的约简集可以正确分类 7 条记录，文献[7]得到的约简集只正确分类 6 条记录。由这个例子也可以看出本文算法得到的约简集更好。

5　实验分析

在决策粗糙集的启发式约简中，按照决策重要度逐渐往约简集 R 中增加属性，避免了属性的组合爆炸问题，极大地减少了计算量；并且在删除冗余属性时，总共只进行了 $|R|$ 次比较来剔除冗余属性，而不是检查 R 的所有子集，因为 R 的非空子集有 $2^{|R|}$ 个，如果 $|R|$ 大，那个针对非空子集的计算量也是很大的。本文算法在寻找一个较大的约简集 R 以及删除冗余属性的过程中，都避免了对超出最小属性约简个数的随机组合情况的搜索。所以算法的求解速度较快，能够得到极小约简集。

为了充分验证该算法的有效性，选取了 UCI 数据库中的 4 个数据集 Dermatology，Zoo，Mushroom，Soybean 来进行实验，令阈值 α 为 0.8。使用该文中提到的决策重要度为启发式函数和文献[7]中提到的属性重要度为启发函数的算法相比较，此处把使用决策重要度的启发式约简算法称为算法 1，把使用属性重要度的启发式约简算法称为算法 2，得出比较结果如表 2 和表 3 所列(由于属性已经按照重要度排序，因此算法进行属性搜索的顺序就固定了，故对于同一个数据集，从理论上来说无论运行多少次，得到的结果都是固定的。对这些数据进行过超过 50 次实验，同一个数据集约简后的结果都不变)。

表 2　不同数据集的约简结果

数据集	样本数	属性数	约简集属性数	
			算法 1	算法 2
Dermatology	358	35	3	5
Zoo	101	17	4	4
Mushroom	8124	23	3	4
Soybean	307	36	2	4

表 3　不同算法的时间比较

数据集	样本数	时间(s)	
		算法 1	算法 2
Dermatology	358	0.039	0.065
Zoo	101	0.017	0.035
Mushroom	8124	0.681	1.194
Soybean	307	0.016	0.051

对表 2 进行分析，可以发现：

(1)对于同一组数据，本文中使用决策重要度为启发函数进行属性约简的算法 1 得到的约简集中的属性个数要少于算法 2，说明本文中启发算法能够得到较小的约简集；

(2)对于不同的数据集，无论样本个数的大小，属性个数的多少，本文算法都能得到较理想的约简结果。

对表 3 进行分析，可知：

(1)对于同一个数据集，算法 1 所用的时间少于算法 2；对于不同的数据集，算法 1 所用的时间都比较少。

(2)两个算法在求等价类时的时间复杂度一样；采取的启发算法不同，所以本文算法 1 在局部求取约简 R 时，时间复杂度较小一点。两个算法的运行环境完全一样，经过多次实验结果的比较，本文算法在速度上有优势，数据集越大优势越明显。

下面将对所有对象的正决策的单调性给出数据实验分析。数据使用上面的 Dermatology，Zoo，Mushroom，令阈值 α

为 0.8，结果如图 1—图 3 所示。

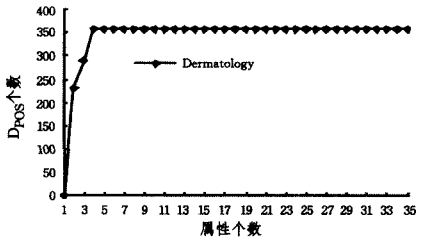

图 1 数据集 Dermatology 不同属性个数时正决策对象个数

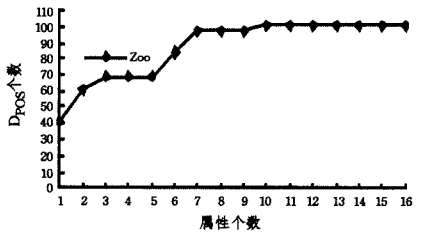

图 2 数据集 Zoo 不同属性个数时正决策对象个数

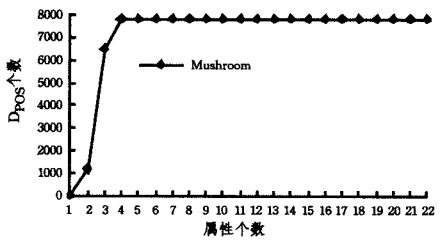

图 3 数据集 Mushroom 不同属性个数时正决策对象个数

由图 1—图 3 可知，正决策对象个数随属性个数的增加逐渐增大(非减)，具有单调性特征。保持所有对象正决策类不变的属性约简方法相当于将所有概率正区域看成 Pawlak 模型的正区域，其他两个概率区域看成边界域，使约简具有稳定性。

综合上述分析可知，本文算法得到的约简集较小，求得一个约简集所花费的时间较少，所以本文算法的约简效果较好，且对于大数据集也比较适合。

结束语 本文深入分析了决策粗糙集的属性约简问题，首先提出新的启发函数，对属性的重要性给出度量；然后基于此启发函数设计了新的决策粗糙集属性约简算法。本文算法的优点是能够得到一个完备的极小的约简集，即约简集中不存在冗余属性；另外算法的时间复杂度低，能够显著提高求解约简集的效率，在针对大规模数据和高维数据约简方面能够得到较好的效果，有很强的实用性。通过 UCI 标准数据集实验验证表明，本文算法具有较好的约简效果，但还需要进一步研究算法对大规模非完备数据的处理能力。

参 考 文 献

[1] Yao Y, Zhao Y. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178(17): 3356-3373

[2] Miao D Q, Hu G R. A heuristic algorithm for reduction of knowledge[J]. Journal of Computer Research and Development, 1999, 36(6): 681-684(in Chinese)
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

[3] Shen W, Zhao J B. A New Heuristic Reduction Algorithm of Rough Sets Decision-Making Table[J]. Computer Technology and Development, 2010, 20(10): 16-20(in Chinese)
沈玮，赵佳宝. 一种新的启发式粗集决策表属性约简算法[J]. 计算机技术与发展，2010，20(10)：16-20

[4] Shi H J, Qin C, Chen H J, et al. Heuristic algorithm of attribute reduction in condition entropy[J]. Computer Engineering and Design, 2008, 29(19): 5014-5015(in Chinese)
施化吉，秦川，陈海军，等. 基于粗糙集的启发式属性约简算法[J]. 计算机工程与设计，2008，29(19)：5014-5015

[5] Liang Y, He Z S. A Novel Feature Selection Heuristic Algorithm Based on Rough Set Theory[J]. Computer Science, 2007, 34(6): 162-165(in Chinese)
梁琰，何中市. 一种基于粗糙集启发式的特征选择算法[J]. 计算机科学，2007，34(6)：162-165

[6] Zhang Y. The Research on Efficient Heuristic Attribute Reduction Algorithm Based on Rough Set[D]. Changsha: Central South University of Forestry and Technology, 2013(in Chinese)
张燕. 基于粗糙集的启发式高效属性约简算法的研究[D]. 长沙：中南林业科技大学，2013

[7] Li H X, Zhou X Z, Li T R, et al. Decision Rough set Theory and its Research Progress[M]. Beijing: Science Press, 2011: 83-89(in Chinese)
李华雄，周献中，李天瑞，等. 决策粗糙集理论及其研究进展[M]. 北京：科学出版社，2011：83-89

[8] Skowron A, Rauszer C. The discernibility matrices and functions in information systems[M]// Intelligent Decision Support. Springer Netherlands, 1992: 331-362

[9] Pawlak Z. Rough sets: Theoretical aspects of reasoning about data[M]. Springer Science & Business Media, 1991

[10] Qian J, Lv P, Yue X D. Research on Attribute Reduction Algorithm and Attribute Core in Decision-Theoretic Rough Set[J]. Journal of Frontiers of Computer Science and Technology, 2014, 8(3): 343-351(in Chinese)
钱进，吕萍，岳晓冬. 决策粗糙集属性约简算法与属性核研究[J]. 计算机科学与探索，2014，8(3)：345-351

[11] Wang G, Yu H, Li T. Decision region distribution preservation reduction in decision-theoretic rough set model[J]. Information Sciences, 2014, 278: 614-640

[12] Liu S H, Sheng Q J, Wu B, et al. Research on efficient algorithms for rough set methods[J]. Chinese Journal of Cmputers-Chinese Edition, 2003, 26(5): 524-529(in Chinese)
刘少辉，盛秋戬，吴斌，等. Rough 集高效算法的研究[J]. 计算机学报，2003，26(5)：524-529

[13] Zhao Y, Wong S K M, Yao Y Y. A note on attribute reduction in the decision-theoretic rough set model[M]// Transactions on rough sets XIII. Springer Berlin Heidelberg, 2011: 260-275

[14] Yao Y Y, Zhao Y, Wang J. On Reduct Construction Algorithms, Rough Sets and Knowledge Technology[C]// Proceedings of the First International Conference of Rough Set and Knowledge Technology (RSKT 2006). 2006: 297-304