文章编号：2096-1618(2017)01-0086-04

基于模糊贴近度的属性约简

卞广钱， 周 磊
(成都信息工程大学应用数学学院，四川 成都 610225)

摘要：属性约简是信息系统中知识发现的重要内容. 基于模糊决策信息系统探讨模糊环境下的属性约简问题. 首先根据模糊贴近度定义模糊决策信息系统中的决策协调集与决策约简集，由此引入一种关于贴近度的属性辨识集，然后结合属性的特征讨论了模糊决策信息系统的属性约简理论，提出一种基于模糊贴近度的属性约简方法，并举例验证和说明此方法的运用。

关 键 词：应用数学；模糊粗糙集；属性约简；辨识集；模糊贴近度；模糊决策信息系统
中图分类号：O159 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2017.01.015

0 引言

20世纪80年代Pawlak[1]提出粗糙集理论，用来研究不确定性问题[2]. 经过几十年的发展，经典的粗糙集模型经过众多专家学者的努力已经得到大规模推广，例如模糊粗糙集[3-4]、概率粗糙集模型[5]、直觉模糊粗糙集、区间值模糊粗糙集等而粗糙及理论也不断渗透到其他相关领域，目前已经应用于医学、金融、专家系统、知识发现[6]、数据挖掘[7]、决策支持系统[8]、机器学习等方面。

属性约简作为粗糙集研究的热点问题，一直以来受到众多研究者的广泛关注。属性约简的基本思想是在保持知识库分类能力不变的前提下，删除其中的冗余或不必要的属性从而达到对知识库简化的作用. 通常来说对知识库中属性的约简并不是唯一的，然而求解最小属性约简和全部的属性约简是N-P难题[9]，因此很多关于属性约简算法的研究中都采用启发式算法. 为更好地适应实际问题的处理，粗糙集理论被推广到模糊信息系统中，而研究模糊环境下的属性约简方法成为不可避免的问题。目前关于这方面的研究主要集中在：朱丽等[10]不完备犹豫模糊混合信息系统的属性约简，谢海[11]不一致不完备直觉模糊信息系统的属性约简，廖毅强等[12]信息系统属性约简的快速算法。郑建兴等[13]一种模糊决策信息系统的属性约简方法，郭庆[14]直觉模糊集信息系统属性约简算法，鲍忠奎[15]直觉模糊目标信息系统的正域约简。田建伟等[16]运用模糊粗糙集的属性约简理论构建了能耗基线模型。邓大勇等[17]研究了F-模糊粗糙集及其约简。目前随着大数据的兴起急需要高效快速的属性约简方法，如何在分布式计算的基础上研究属性约简方法也是重要的研究方向。定义了模糊决策信息系统中基于模糊贴近度的属性辨识集，在此基础上研究了一类模糊决策信息系统的属性约简问题。

收稿日期：2016-05-05
基金项目：国家自然科学基金资助项目(11226262)

1 模糊集和模糊贴近度

所讨论的论域U都是非空有限集.

定义1 设U是论域，$A:U\to[0,1]$，则称A是U上的一个模糊集合. $A(x)$称为模糊集合A的隶属函数，对$\forall x\in U$，$A(x)$表示x隶属于模糊集合A的程度，简称为隶属度. U上全体模糊集合所组成的类称为U的模糊幂集，用$F(U)$表示. $A\in F(U)$意指A是U上的一个模糊集合.

一般地，一个模糊集A可以表示为$A=\{(x,A(x))\mid x\in U\}$. 当$A$为有限集时可以表示为$A=\sum A(x_i)/x_i$，如果$A$为无限集，则可以表示为$A=\int A(x)/x$，特别的，$U=\{(x,1)\mid x\in U\}$，$\varnothing=\{(x,0)\mid x\in U\}$.

定义2 设$A,B,C\in F(U)$，$N:F(U)\times F(U)\to[0,1]$称为模糊贴近度，当满足：

(1) $N(A,B)=N(B,A)$；

(2) $N(A,A)=1$，$N(U,\varnothing)=0$；

(3) 如果$A\subseteq B\subseteq C$，则$N(A,C)\le N(A,B)\wedge N(B,C)$.

根据贴进度的定义，$\forall A,B,C,D\in F(U)$，可以得到以下基本性质：

(1) $N(A,A^c)=0$；

(2) $N(A,F)=N(A^c,F)$;

(3) $N(A,B)=N(A^c,B^c)$;

(4) $N(A,B)=1\Rightarrow A=B$;

(5) $N(A,D)\leq N(B,C)$.

常见的模糊贴近度有:

海明贴近度

$$N_H(A,B)=1-\frac{1}{n}\sum_{i}^{n}|A(x_i)-B(x_i)|;$$

欧几里得贴近度

$$N_E(A,B)=1-\frac{1}{\sqrt{n}}[\sum_{i=1}^{n}(A(x_i)-B(x_i))^2]^{\frac{1}{2}};$$

格贴近度

$$N_L(A,B)=(A\circ B)\wedge(\overset{\wedge}{A\circ B})^c.$$

式中当 U 为有限论域时,$A\circ B=\bigvee_{i=1}^{n}(A(u_i)\wedge B(u_i))$.

式中当 U 为无限论域时,$A\circ B=\bigvee_{u\in U}(A(u)\wedge B(u))$.

下文讨论的模糊属性约简中,采用的是海明贴近度.

2　基于模糊贴近度的属性约简

定义 3　设 $U=\{x_1,x_2,\cdots,x_n\}$ 为论域, $A=\{a_1,a_2,\cdots,a_m\}$ 为属性集, $F=\{f_l:U\to V(l\leq m)\}$,其中 V 是有限区间或数集, $D=\{D_j:U\to[0,1]\}(j\leq r)$ 是模糊决策集族,称 $G=(U,A,F,D)$ 为模糊决策信息系统.

对模糊决策信息系统 $G=(U,A,F,D)$,记 $R_B=\{(x_i,x_j)\mid f_l(x_i)=f_l(x_j),\forall a_l\in B\},(B\subseteq A)$ 显然, R_B 是 U 上的等价关系, x_i 的等价类为 $[x_i]_B=\{x_j\mid(x_i,x_j)\in R_B\}$. 则

$$R_a=\{(x_i,x_j)\mid f_a(x_i)=f_a(x_j),a\in A\}.$$

设 $G=(U,A,F,D)$ 是模糊决策信息系统, N 是 U 上的模糊贴近度,记

$N_B(D_k)(x)=N(D_k,[x]_B)$;

$\sigma_B(x)=(N_B(D_1)(x),N_B(D_2)(x),\cdots,N_B(D_r)(x))$;

$\sigma_B=\{(x_i,x_j)\mid\sigma_B(x_i)=\sigma_B(x_j)\}$;

定义 4　设 $G=(U,A,F,D)$ 为模糊决策信息系统, $B\subseteq A$,如果 $\sigma_B(x)=\sigma_A(x)(\forall x\in U)$,则称 B 为 G 的决策协调集,如果对任意的 $C\subseteq B(C\neq B)$, C 不是 G 的决策协调集,则称 B 为 G 的决策约简集.

定理 1　设 $G=(U,A,F,D)$ 为模糊决策信息系统, $B\subseteq A$,则 $R_B\subseteq\sigma_A\Leftrightarrow B$ 是 G 的决策约简集.

证明 "$\Leftarrow$" 对任意的 $x_i\in U$,假设 $\sigma_B(x_i)=\sigma_A(x_i)$, $\forall(x_i,x_j)\in R_B$, $[x_i]_B=[x_j]_B$ 则 $N(D_k,[x_i]_B)=N(D_k,[x_j]_B)(k\leq r)$. 所以 $\sigma_B(x_i)=\sigma_B(x_j)$ 即 $\sigma_B(x_i)=\sigma_A(x_j)$,所以 $(x_i,x_j)\in\sigma_A$ 故 $R_B\subseteq\sigma_A$.

"$\Rightarrow$" ,假设 $R_B\subseteq\sigma_A$,则 $\forall(x_i,x_j)\in R_B$, $(x_i,x_j)\in\sigma_A$. 同样当 $[x_i]_B=[x_j]_B$ 时,有 $\sigma_A(x_i)=\sigma_A(x_j)$,由此 $N(D_k,[x_i]_A)=N(D_k,[x_j]_A)(k\leq r)$.

同时, $\forall x_i,x_j\in[x_i]_B$, $[x_i]_B=\cup\{[x_j]_A\mid x_j\in[x_i]_B\}$. 当 $[x_i]_A\neq[x_j]_A$ 时,有 $[x_i]_A\cap[x_j]_A=\varnothing$. 所以 $[x_i]_B=[x_j]_A$,于是 $[x_i]_B=[x_i]_A$. 另外结合贴近度定义可知, $N(D_k,[x_i]_B)=N(D_k,[x_i]_A)$,因此 $\sigma_B(x_i)=\sigma_A(x_i)$,即 B 是 G 的决策约简集.

定义 5　设 $G=(U,A,F,D)$ 为模糊决策信息系统,记

$$S_{\sigma_A}(x_i,x_j)=\begin{cases}\{a\in A\mid(x_i,x_j)\notin R_a\},&\sigma_A(x_i)\neq\sigma_A(x_j);\\ \varnothing,&\sigma_A(x_i)=\sigma_A(x_j).\end{cases}$$

称 $S_{\sigma_A}(x_i,x_j)$ 为 x_i , x_j 关于 σ_A 的属性辨识集.

定理 2　设 $G=(U,A,F,D)$ 为模糊决策信息系统, $B\subseteq A$, B 是 G 的决策约简集,则对 $\forall x_i,x_j\in U$,当 $\sigma_A(x_i)\neq\sigma_A(x_j)$ 时, $B\cap S_{\sigma_A}(x_i,x_j)\neq\varnothing$.

证明　B 是 G 的决策约简集 $\Leftrightarrow$ 当 $\sigma_A(x_i)\neq\sigma_A(x_j)$ 时, $[x_i]_B\cap[x_j]_B=\varnothing\Leftrightarrow$ 当 $\sigma_A(x_i)\neq\sigma_A(x_j)$ 时存在 $a\in B$ 使得 $(x_i,x_j)\notin R_a\Leftrightarrow$ 存在 $a\in B$,满足 $a\in S_{\sigma_A}(x_i,x_j)\Leftrightarrow$ 当 $\sigma_A(x_i)\neq\sigma_A(x_j)$ 时, $B\cap S_{\sigma_A}(x_i,x_j)\neq\varnothing$.

定义 6　设 $G=(U,A,F,D)$ 为模糊决策信息系统, $RED(G)$ 是 G 的全体决策约简集,称 $CORE(G)=\cap RED(G)$ 为 G 的核,其中的元素称为核属性. $A-\cup RED(G)$ 称为 G 的不必要属性集,其中的元素称为不必要属性. $\cup RED(G)-CORE(G)$ 称为 G 的相对必要属性集,其中的元素称为相对必要属性.

定理 3　设 $G=(U,A,F,D)$ 为模糊决策信息系统,则以下说法等价:

(1) a 是 G 的核属性;

(2) $\exists x_i,x_j\in U$,满足 $S_{\sigma_A}(x_i,x_j)=\{a\}$;

(3) $R_{A-\{a\}}\not\subset\sigma_A$.

证明　"(1)$\Rightarrow$(2)",假设每个辨识集至少含有两个属性,其中一个为 a ,记 $B=\bigcup_{i,j}(S_{\sigma_A}(x_i,x_j)-\{a\})$. 当 $\sigma_A(x_i)\neq\sigma_A(x_j)$ 时,必定有 $B\cap S_{\sigma_A}(x_i,x_j)\neq\varnothing$,由定理 1 知 B 是 G 的决策协调集且 $a\notin B$. 因此,存在一个决策约简集 $B'\subseteq B$,且 $a\notin B'$,这与 a 是 G 的核属性矛盾.

"(2)$\Rightarrow$(3)",如果 $\exists x_i,x_j\in U$,满足 $S_{\sigma_A}(x_i,x_j)=\{a\}$,由定义 5 可知 $\sigma_A(x_i)\neq\sigma_A(x_j)$,即 $(x_i,x_j)\notin\sigma_a$,同时 $(x_i,x_j)\notin R_a$,又因为 $(x_i,x_j)\in R_b(\forall b\in A-\{a\})$,则 $(x_i,x_j)\in R_{A-\{a\}}$,所以 $R_{A-\{a\}}\not\subset\sigma_A$.

"(3)⇒(1)",假设 a 不是 G 的核,则必存在一个约简集 $B \subseteq A$ 满足 $a \notin B$,即 $B \subseteq A - \{a\}$,因此 $R_{A-\{a\}} \subseteq R_B$. 又由定理1可知, $R_B \subseteq \sigma_A$,从而 $R_{A-\{a\}} \subseteq \sigma_A$,与已知矛盾,所以 a 是 G 的核属性.

定理4 设 $G=(U,A,F,D)$ 为模糊决策信息系统, a 是 G 的不必要属性 $\Leftrightarrow R(a) \subseteq \sigma_A \cup R_a$,其中 $R(a)=\cup \{R_{B-\{a\}} \mid R_B \subseteq \sigma_A, B \subseteq A\}$.

证明 "⇒"假设 a 是 G 的不必要属性,则 a 不属于任何一个约简集,所以对每个 $R_B \subseteq \sigma_A (B \subseteq A)$,都有 $R_{B-\{a\}} \subseteq \sigma_A$. 另外若 $R_{B-\{a\}} \not\subseteq \sigma_A$,则对于 $B' \subseteq B-\{a\}$,有 $R_{B-\{a\}} \subseteq R_{B'}$, $R_{B'} \not\subseteq \sigma_A$,因此 B 是 G 的决策约简集且 $a \in B$,这与已知矛盾,故 $\forall B \subseteq A$,当 $R_B \subseteq \sigma_A$ 时,必然有 $R_{B-\{a\}} \subseteq \sigma_A \cup R_a$.

"⇐"假设 $R(a) \subseteq \sigma_A \cup R_a$, $\forall B \subseteq A$,当 $R_B \subseteq \sigma_A$ 时,有 $R_{B-\{a\}} \subseteq \sigma_A \cup R_a$,则 $R_{B-\{a\}} \cap R_a^C \subseteq \sigma_A$. 因此, $R_{B-\{a\}}=R_B \cup (R_{B-\{a\}} \cap R_a^C) \subseteq \sigma_A$,所以 a 不在任何一个决策约简集中,即 a 是 G 的不必要属性.

例1 设 $G=(U,A,F,D)$ 为模糊决策信息系统,如表1所示。

表1 模糊决策信息系统

U	a_1	a_2	a_3	D_1	D_2	D_3
x_1	2	1	3	0.9	0.2	0.5
x_2	3	2	1	0.5	0.7	0.4
x_3	2	1	3	0.8	0.4	0.2
x_4	2	2	3	0.2	0.8	0.3
x_5	1	1	4	0.1	0.3	0.9
x_6	1	1	2	0.2	0.5	1.0
x_7	3	2	1	0.4	1.0	0.4
x_8	1	1	4	0.2	0.4	0.6
x_9	2	1	3	0.6	0.3	0.3
x_{10}	3	2	1	0.1	0.9	0.2

条件属性集 A 将论域 U 分为以下5类:

$[x_1]_A=\{x_1,x_3,x_9\}$, $[x_2]_A=\{x_2,x_7,x_{10}\}$, $[x_4]_A=\{x_4\}$, $[x_5]_A=\{x_5,x_8\}$, $[x_6]_A=\{x_6\}$.

利用模模糊贴近度计算得到:

$N_A(D_1)(x_1)=0.76$, $N_A(D_1)(x_2)=0.5$, $N_A(D_1)(x_4)=0.54$, $N_A(D_1)(x_5)=0.46$, $N_A(D_1)(x_6)=0.54$, $N_A(D_2)(x_1)=0.33$, $N_A(D_2)(x_2)=0.66$, $N_A(D_2)(x_4)=0.51$, $N_A(D_2)(x_5)=0.39$, $N_A(D_2)(x_6)=0.45$,

$N_A(D_3)(x_1)=0.32$, $N_A(D_3)(x_2)=0.42$, $N_A(D_3)(x_4)=0.49$, $N_A(D_3)(x_5)=0.62$, $N_A(D_3)(x_6)=0.62$.

由此可得:

$\sigma_A(x_1)=\sigma_A(x_3)=\sigma_A(x_9)=(0.76,0.33,0.32)$,

$\sigma_A(x_2)=\sigma_A(x_7)=\sigma_A(x_{10})=(0.5,0.66,0.42)$,

$\sigma_A(x_4)=(0.54,0.51,0.49)$,

$\sigma_A(x_5)=\sigma_A(x_8)=(0.46,0.39,0.62)$,

$\sigma_A(x_6)=(0.54,0.45,0.62)$.

取 $B_1=\{a_1,a_2\}$, $B_2=\{a_1,a_3\}$, $B_3=\{a_2,a_3\}$.

从而

$[x_1]_{B_1}=\{x_1,x_3,x_9\}$, $[x_2]_{B_1}=\{x_2,x_7,x_{10}\}$, $[x_4]_{B_1}=\{x_4\}$, $[x_5]_{B_1}=\{x_5,x_6,x_8\}$, $[x_1]_{B_2}=\{x_1,x_3,x_4,x_9\}$,

$[x_2]_{B_2}=\{x_2,x_7,x_{10}\}$, $[x_5]_{B_2}=\{x_5,x_8\}$, $[x_6]_{B_2}=\{x_6\}$,

$[x_1]_{B_3}=\{x_1,x_3,x_9\}$, $[x_2]_{B_3}=\{x_2,x_7,x_{10}\}$, $[x_4]_{B_3}=\{x_4\}$, $[x_5]_{B_3}=\{x_5,x_8\}$, $[x_6]_{B_3}=\{x_6\}$.

同样可得:

$\sigma_{B_1}(x_1)=\sigma_{B_1}(x_3)=\sigma_{B_1}(x_9)=(0.76,0.33,0.32)$,

$\sigma_{B_1}(x_2)=\sigma_{B_1}(x_7)=\sigma_{B_1}(x_{10})=(0.5,0.66,0.42)$,

$\sigma_{B_1}(x_4)=(0.5,0.51,0.49)$,

$\sigma_{B_1}(x_5)=\sigma_{B_1}(x_6)=\sigma_{B_1}(x_8)=(0.4,0.39,0.62)$,

$\sigma_{B_2}(x_1)=\sigma_{B_2}(x_3)=\sigma_{B_2}(x_4)=\sigma_{B_2}(x_9)=(0.7,0.33,0.32)$,

$\sigma_{B_2}(x_2)=\sigma_{B_2}(x_7)=\sigma_{B_2}(x_{10})=(0.5,0.66,0.42)$,

$\sigma_{B_2}(x_5)=\sigma_{B_2}(x_8)=(0.46,0.39,0.62)$,

$\sigma_{B_2}(x_6)=(0.54,0.45,0.62)$,

$\sigma_{B_3}(x_1)=\sigma_{B_3}(x_3)=\sigma_{B_3}(x_9)=(0.76,0.33,0.32)$,

$\sigma_{B_3}(x_2)=\sigma_{B_3}(x_7)=\sigma_{B_3}(x_{10})=(0.5,0.66,0.42)$,

$\sigma_{B_3}(x_4)=(0.54,0.51,0.49)$,

$\sigma_{B_3}(x_5)=\sigma_{B_3}(x_8)=(0.46,0.39,0.62)$,

$\sigma_{B_3}(x_6)=(0.54,0.45,0.62)$.

因此, $\forall x_i \in U$,有 $\sigma_A(x_i)=\sigma_{B_3}(x_i)$,但是 $\sigma_A(x_5) \neq \sigma_{B_1}(x_5)$, $\sigma_A(x_1) \neq \sigma_{B_2}(x_1)$,从而由定义4可知, $B_3=\{a_2,a_3\}$ 为 G 的协调集且为约简集; B_1、B_2 不是 G 的协调集,此结论同样可以由定理2得到.

3 结束语

通过模糊贴近度定义决策协调集与决策约简集,由此引入属性辨识集的概念,根据辨识集的思想,结合属性的特征,讨论一种基于贴近度属性约简方法,为研究模糊决策信息系统的属性约简问题提供一种新的思路,下一步工作中我们将结合模糊决策方法定义继续深入研究一般模糊信息系统中的属性约简问题.

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences ,1982,11:341-356.

[2] Pawlak Z. Rough sets: Theoretieal Aspeets of Reasoning about Data [M]. Kluwer Aeademie

Publishers:Boston. 1991:66-90.
[3] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets [J]. International Journal of General System, 1990, 17:191-208.
[4] Dubois D, Prade H. Putting rough sets and fuzzy sets together[C]. Slowinski R, Intelligent Decision Support . [S. l.]: Kluwer Academic , Dordrecht, 1992:203-232.
[5] 张文修. 粗糙集理论与方法[M]. 北京:科学出版社,2001.
[6] Chmielew ski M R, Grzymala-Busse J W. Global discretization of continuous attributes as preprocessing for machine learning[J]. International Journal of Approximate Reasoning, 1996, 15:319-331.
[7] Chan C C. A rough set approach to attribute generalization in datamining [J]. Journal of Information Sciences, 1998, 107:169-176.
[8] McSherry D. Knowledge discovery by inspection [J]. Decision Support Systems, 1997, 21:43-47.
[9] Wong S K M, Ziarko W. Optimal decision rules in decision table [J]. Bulletin of Polish Academy of sciences, 1985, 33(11-12):693-696.
[10] 朱丽,朱传喜,张小芝. 不完备犹豫模糊混合信息系统的属性约简[J]. 模糊系统与数学, 2015,29(2):150-156.
[11] 谢海. 不一致不完备直觉模糊信息系统的属性约简[J]. 数学的实践与认识,2015,45(21): 267-273.
[12] 廖毅强, 桂现才. 信息系统属性约简的快速算法[J]. 计算机工程与设计,2008,29(18):4804-4806.
[13] 郑建兴,李德玉. 一种模糊决策信息系统的属性约简方法[J]. 中北大学学报(自然科学版), 2011,32(1):115-118.
[14] 郭庆,杨善林,刘文军. 直觉模糊集信息系统属性约简算法[J]. 模糊系统与数学,2014,28(4):138-143.
[15] 鲍忠奎,杨善林. 直觉模糊目标信息系统的正域约简[J]. 中国科学技术大学学报,2015,45(4):329-336.
[16] 田建伟;李志忠;陈海红. 基于模糊粗糙集属性约简理论的能耗基线模型[J]. 太阳能学报, 2015,36(10):2347-2353.
[17] 邓大勇;徐小玉;裴明华. F-模糊粗糙集及其约简[J]. 浙江师范大学学报(自然科学版), 2015,38(1):58-66.

Attribute Reduction based on Fuzzy Closeness Degree

BIAN Guang-qian, ZHOU Lei
(College of Applied Mathematics, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Attribute reduction is an important content of knowledge discovery in information systems. This paper first introduces the notion and properties of fuzzy nearness degree. Then, According to the fuzzy nearness the decision coordination set and the decision reduction set for the fuzzy decision information system are defined. Furthermore, the attribute discernibility set is established by the fuzzy nearness degree, and meanwhile, the attribute reduction of fuzzy decision information system is discussed. As a result, a new attempt of attribute reduction theories and methods based on fuzzy nearness degree is proposed in fuzzy information systems, and an example is given to illustrate the application of this method.

Keywords: applied mathematics; fuzzy rough sets; attribute reduction; discernibility attribute set; fuzzy closeness degree; fuzzy decision information system