维普资讯 http://www.cqvip.com

基于粗糙集的形式背景属性约简及属性特征*)

李同军[1,2]　张文修[1]　马建敏[1]
(西安交通大学理学院信息与系统科学研究所　西安 710049)[1]
(浙江海洋学院信息学院　浙江 舟山 316004)[2]

摘　要　本文在形式背景中,定义了一种上、下近似算子,给出近似算子的性质。基于算子给出形式背景属性约简的定义,得到属性约简的判定理论,刻画出不同类型属性的特征。最后,给出形式背景属性约简的方法。

关键词　粗糙集,形式背景,属性约简,属性特征

Attribute Reductions and Attribute Features of Formal Contexts Based on a Type of Rough Sets

LI Tong-Jun[1,2]　ZHANG Wen-Xiu[1]　MA Jian-Min[1]
(Faculty of Science, Institute for Information and System Science, Xi'an Jiaotong University, Xi'an 710049)[1]
(Information College, Zhejiang Ocean University, Zhoushan, Zhejiang 316004)[2]

Abstract　Rough sets theory is a kind of tools for data analysis and knowledge discovery. The knowledge hidden in universe or information systems can be revealed by using the lower and upper approximation operators. Representing knowledge with more little attributes is an interesting subject. In this paper, a type of lower and upper approximation is defined in formal contexts— pecial information systems, and the properties of the operators are investigated. Based on the defined approximation operators, the definition of attribute reduction is given, and the theory and method of attribute reduction is examined. The given feature characterization of attributes shows the relationships between each attribute and all attribute reductions.

Keywords　Rough sets, Formal contexts, Attribute reductions, Attribute features

1　引言

粗糙集理论是波兰数学家 Pawlak[1] 于 1982 年提出的一种数据分析理论。利用上、下近似可以近似地表示不精确或不确定的知识,进而可以挖掘出隐藏在信息表中的确定性知识。Pawlak 粗糙集依据论域上的等价关系来定义近似算子,等价关系是满足自反性、对称性和传递性的二元关系。在实际问题中,许多情况下论域的知识不是由等价关系提供的,可能是由一种稍弱的二元关系或覆盖提供的,因此有必要对 Pawlak 粗糙集模型进行推广。文[2,3,7]在这方面做了许多工作,使近似算子可以定义在一般的二元关系和一般的论域覆盖之上。

属性的约简是粗糙集理论和数据挖掘中的核心问题之一,它是简化知识表示和获取实用规则的前提,在这方面已有了相当丰富的成果[4~6]。本文就是在一种特殊的信息系统——形式背景中引入上、下近似运算,给出保证近似算子不变的属性约简定义,进而阐述相应的属性约简理论和方法。本文的工作对一般信息系统的知识发现研究有一定的意义。

2　形式背景与粗糙集

定义 2.1　称(U,A,I)是形式背景。其中$U=\{x_1,x_2,\cdots,x_n\}$为对象集,每个$x_i(i\leqslant n)$称为一个对象;$A=\{a_1,a_2,\cdots,a_m\}$为属性集,每个$a_j(j\leqslant m)$称为一个属性;I是U和A之间的二元关系,$I\subseteq U\times A$。若$(x,a)\in I$,则称x具有属性a,记为xIa。对于$B\subseteq A$,称(U,B,I_B)为(U,A,I)的子形式背景,其中$I_B=I\cap(U\times B)$。

显然$I_A=I$。

设(U,A,I)是一个形式背景,$B\subseteq A$,对于每个$x\in U$,记$f_B(x)=\{a\in B:xI_Ba\}$,其实$f_B(x)=f_A(x)\cap B$,称$f_B(x)$为对象x在B中属性集;对于每个$a\in B$,记$g(a)=\{x\in U:xIa\}$,称$g(a)$为属性a的对象集。若对于每个$x\in U$,$f_B(x)\neq\phi$,$f_B(x)\neq A$,且对于每个$a\in B$,$g(a)\neq\phi$,$g(a)\neq U$,则称形式背景(U,B,I_B)是正则的。以下假设(U,A,I)是正则的。

对于形式背景(U,B,I_B),$\{g(a):a\in B\}$是对象集U上的一个子集类。如果(U,B,I_B)是正则的,那么$\{g(a):a\in B\}$是U上的一个覆盖。根据文[3]中的方法,基于$\{g(a):a\in B\}$可以定义上、下近似如下。

定义 2.2　设(U,B,I_B)为(U,A,I)的子形式背景,对于$X\subseteq U$,令

$$\overline{apr}_B(X)=\{x\in U:\exists a(x\in g(a),g(a)\cap X\neq\phi)\},$$

$$\underline{apr}_B(X)=\{x\in U:\forall a(x\in g(a)\Rightarrow g(a)\subseteq X)\},$$

则称$\overline{apr}_B(X)$和$\underline{apr}_B$分别为X在(U,B,I_B)的上、下近似,分别称$\overline{apr}_B$和$\underline{apr}_B:P(U)\to P(U)$为上、下近似算子。

命题 2.1[3]　上、下近似算子$\overline{apr}_B$和$\underline{apr}_B$满足下列性质:$X,Y\subseteq U$,

(1)$\underline{apr}_B(X)=\sim\overline{apr}_B(\sim X)$,$\overline{apr}_B(X)=\sim\underline{apr}_B(\sim X)$;

(2)$\underline{apr}_B(X\cap Y)=\underline{apr}_B(X)\cap\underline{apr}_B(Y)$,$\overline{apr}_B(X\cup Y)=$

*)国家自然科学基金(60373078)、国家"九七三"重点基础研究发展规划资助项目(2002CB312206)。**李同军**　博士生,副教授,主要研究方向为粗糙集、概念格;**张文修**　教授,博士生导师,主要研究方向为粗糙集、模糊集等;**马建敏**　博士生,主要研究方向为人工智能、粒计算。

维普资讯 http://www.cqvip.com

$\overline{apr}_B(X)\cup\overline{apr}_B(Y)$；

(3)若 $X\subseteq Y$，则$\underline{apr}_B(X)\subseteq\underline{apr}_B(Y)$，$\overline{apr}_B(X)\subseteq\overline{apr}_B(Y)$；

(4)$\underline{apr}_B(X)\subseteq\overline{apr}_B(X)$；

(5)若 $B_1\subseteq B_2\subseteq A$，则$\underline{apr}_{B_2}(X)\subseteq\underline{apr}_{B_1}(X)\subseteq\overline{apr}_{B_1}(X)\subseteq\overline{apr}_{B_2}(X)$；

(6)若 $B_1,B_2\subseteq A$，则$\underline{apr}_{B_1\cup B_2}(X)=\underline{apr}_{B_1}(X)\cap\underline{apr}_{B_2}(X)$，$\overline{apr}_{B_1\cup B_2}(X)=\overline{apr}_{B_1}(X)\cup\overline{apr}_{B_2}(X)$。

对于任意的 $x\in U$，若令 $n_B(x)=\bigcup_{a\in f_B(x)}g(a)$，则$\{n_B(x):x\in U\}$是 U 上的一个邻域系统。对于 $a\in A,x\in U$，若 xIa，则 $n_{\{a\}}(x)=g(a)$，否则 $n_{\{a\}}(x)=\phi$。

命题 2.2 对于任意的 $x\in U$，有：

(1)$x\in n_B(x)$；

(2)$y\in n_B(x)\Leftrightarrow x\in n_B(y)\Leftrightarrow f_B(x)\cap f_B(y)\neq\phi$；

(3)若 $B_1\subseteq B_2$，则 $n_{B_1}(x)\subseteq n_{B_2}(x)$；

(4)若 $B_1,B_2\subseteq A$，则 $n_{B_1\cup B_2}(x)=n_{B_1}(x)\cup n_{B_2}(x)$；

(5)$\overline{apr}_B(\{x\})=n_B(x)$。

由命题 2.2 的(1)和(2)可知邻域系统$\{n_B(x):x\in U\}$决定了一个相似的二元关系，即满足自反性、对称性的二元关系，$R_B=\{(x,y)\in U\times U:f_B(x)\cap f_B(y)\neq\phi\}$。对于任意的 $x\in U,B\subseteq A$，若令 $R_B(x)=\{y\in U:(x,y)\in R_B\}$，则有：

命题 2.3 (1)$\forall x\in U$，有 $n_B(x)=R_B(x)$；

(2)若 $B_1\subseteq B_2\subseteq A$，则 $R_{B_1}\subseteq R_{B_2}\subseteq R_A$；

(3)若 $B_1,B_2\subseteq A$，则 $R_{B_1\cup B_2}=R_{B_1}\cup R_{B_2}$。

3 形式背景中属性约简及属性特征

定义 3.1 设(U,A,I)是一形式背景，$B\subseteq A$，若$\underline{apr}_B=\underline{apr}_A$(或$\overline{apr}_B=\overline{apr}_A$)，则称 B 是关于近似算子的属性协调集。若 $D\subseteq A$ 是属性协调集，且对于任意的 $a\in D$，$D-\{a\}$不是属性协调集，则称 D 为属性约简集。

显然，若 $D\subseteq A$ 是属性约简集，则有 D 是属性协调集，且$\underline{apr}_{D-\{a\}}\neq\underline{apr}_D$，$\overline{apr}_{D-\{a\}}\neq\overline{apr}_D$。

定理 3.1 形式背景(U,A,I)的属性约简集总是存在的。

形式背景(U,A,I)的属性约简集可能不止一个。

定理 3.2 设(U,A,I)是形式背景，$B\subseteq A$，则下列命题等价：

(1)B 是属性协调集；

(2)$\forall x\in U$，$n_B(x)=n_A(x)$；

(3)$R_B=R_A$；

(4)$R_A\subseteq R_B$；

(5)$|R_B|=|R_A|$。

其中$|\cdot|$表示集合·中元素的个数。

证明：由命题 2.2 的(5)和定义 3.1 以及命题 2.3 易得。

定理 3.3 设(U,A,I)是形式背景，$D\subseteq A$，则 D 是约简集的充要条件是 D 为协调集，且对于任意的 $a\in D$，$\exists x_0\in U$，使得 $n_D(x_0)\neq n_{D-\{a\}}(x_0)$。

证明：由定理 3.2 和定义 3.1 可得。

正则形式背景的属性协调集对应的子形式背景未必是正则的。若(U,A,I)是正则形式背景，$B\subseteq A$ 是属性协调集。形式背景(U,B,I_B)显然满足：对于每个 $a\in B$，$g(a)\neq\phi$，$g(a)\neq U$。同时还满足：对于每个 $x\in U$，$f_B(x)\neq\phi$。事实上，由(U,A,I)的正则性可知，对于每个 $x\in U$，$f_A(x)\neq\phi$，所以 $f_A(x)\cap f_A(x)=f_A(x)\neq\phi$，即 $x\in R_A(x)=n_A(x)$。又因为 B 是属性协调集，根据定理可知 $x\in n_B(x)=n_A(x)$，所以 $\exists a\in B$，$a\in f_B(x)$，即 $f_B(x)\neq\phi$。但是对于某些 $x\in U$，可能有 $f_B(x)=B$。因此，(U,B,I_B)未必是正则的。

在形式背景(U,A,I)中，对于属性集 A 中的属性，可根据其对象集是否相同对它们进行分类。对于 $a\in A$，记$[a]=\{b\in A:g(a)=g(b)\}$，则$\{[a]:a\in A\}$是属性集 A 的一个划分。

定义 3.2 在形式背景(U,A,I)中，$a\in A$，若$\forall x\in U$，都有 $n_{A-[a]}(x)=n_A(x)$，则称$[a]$是可约属性类，a 是类可约属性，否则称$[a]$是不可约属性类，a 是类不可约属性。

定理 3.4 在形式背景(U,A,I)中，$a\in A$，以下命题等价：

(1)$[a]$是可约属性类；

(2)$R_{A-[a]}=R_A$；

(3)$\underline{apr}_{A-[a]}=\underline{apr}_A$；

(4)$\overline{apr}_{A-[a]}=\overline{apr}_A$。

定理 3.5 设 D 是形式背景(U,A,I)的一个属性约简集，则对于任意 $a\in A$，都有$|[a]\cap D|=0$或者$|[a]\cap D|=1$。

证明：假设存在一个属性 $a\in A$，$[a]\cap D\neq\phi$，而且$|[a]\cap D|>1$。不妨设 $a,b\in[a]\cap D$。由于 $a,b\in[a]$，因此 $g(a)=g(b)$。由于属性约简集必是属性协调集，因此对于任意的 $x\in U$，有 $n_A(x)=n_D(x)$。若 $x\in g(a)$，则 $n_{\{a\}}(x)=n_{\{b\}}(x)=g(a)$，$n_A(x)=n_D(x)=n_{D-\{a,b\}}(x)\cup n_{\{a\}}(x)\cup n_{\{b\}}(x)=n_{D-\{a,b\}}(x)\cup g(a)\cup g(b)=n_{D-\{a,b\}}(x)\cup g(a)=n_{D-\{b\}}(x)$，即 $n_A(x)=n_{D-\{b\}}(x)$；若 $x\notin g(a)$，则 $n_{\{a\}}(x)=n_{\{b\}}(x)=\phi$，同样可得 $n_A(x)=n_{D-\{b\}}(x)$。综上，对于任意 $x\in U$，$n_A(x)=n_{D-\{b\}}(x)$。根据定理 3.2 知道 $D-\{b\}$是属性协调集，这与 D 是约简集相矛盾，因此定理得证。

根据属性与各属性约简集的关系，可以对 A 中的属性进行如下的分类。

定义 3.3 设$\{D_i\}(i\in\tau)$是形式背景(U,A,I)的所有属性约简集，把 A 中属性分成 3 类：

(1)核心属性集：$C=\bigcap_{i\in\tau}D_i$，

(2)相对必要属性集：$K=\bigcup_{i\in\tau}D_i-C$，

(3)不必要属性集：$N=A-\bigcup_{i\in\tau}D_i$。

定理 3.6 设(U,A,I)是形式背景，则下列命题等价：

(1)$a\in A$ 是核心属性；

(2)$R_A\not\subset R_{A-\{a\}}$；

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

(4)$[a]$是不可约属性类，且$|[a]|=1$。

证明：(1)⇒(2)：假设 $R_A\subseteq R_{A-\{a\}}$，由定理 3.2 可知 $A-\{a\}$是属性协调集，显然至少存在一个约简集 $D_i\subseteq A-\{a\}$，所以 $a\notin D_i$，这与 a 是核心属性相矛盾。

(2)⇒(1)：假设 $a\in A$ 不是核心属性，则存在一个属性约简集 D_i，$a\notin D_i$。所以 $D_i\subseteq A-\{a\}$，$R_A=R_{D_i}\subseteq R_{A-\{a\}}$，这与 $R_A\not\subset R_{A-\{a\}}$矛盾。

(2)⇒(4)：若$|[a]|>1$，由定理 3.4 的证明可知 $R_A=R_{A-\{a\}}$，这与 $R_A\not\subset R_{A-\{a\}}$矛盾，所以$|[a]|=1$。再由 $R_A\not\subset R_{A-\{a\}}=R_{A-[a]}$可得$[a]$是不可约属性类。

(4)⇒(2)和(2)⇒(3)都是显然的。

定理 3.7 设(U,A,I)是形式背景，则下列命题等价：

(1)$a\in A$ 是不必要属性；

(2)$R_{\{a\}}\subseteq R_C$，$R_{\{a\}}\subseteq R_{A-\{a\}}$。

维普资讯 http://www.cqvip.com

证明：(1)⇒(2)：$a\in A$ 是不必要属性，由定理 3.6 可知 $R_{\{a\}}\subseteq R_{A-\{a\}}$。假设 $R_{\{a\}}\not\subset R_C$，令 $D=C\cup\{a\}$，则 $R_D\not\subset R_C$。若 $R_A\subseteq R_D$，则 D 是一个约简集，且 $a\in D$，这与 $a\in N$ 矛盾。若 $R_A\not\subset R_D$，显然存在某个约简集 $D_i\supseteq D$，这时有 $a\in D\subseteq D_i$，这也与 $a\in N$ 矛盾。

(2)⇒(1)：对于任意约简集 D_i，若 $a\in D_i$，由 $R_{\{a\}}\subseteq R_{A-\{a\}}$ 可知 $a\notin C$，则有 $R_{D_i}=R_C\cup R_{D_i-C\cup\{a\}}\cup R_{\{a\}}=R_C\cup R_{D_i-C\cup\{a\}}=R_{D_i-\{a\}}$，由定理 3.2 知 $D_i-\{a\}$ 是属性协调集，这与 D_i 是约简集矛盾。所以对于任意的约简集 D_i，都有 $a\notin D_i$，即 $a\in N$。

定理 3.8 不必要属性一定是类可约属性。

证明：设 $a\in N$，则由定理 3.7 的(2)可知 $[a]\cap C=\phi$，且 $R_{\{a\}}\subseteq R_C$。因为 $R_{[a]}=R_{\{a\}}$，所以 $R_{[a]}\subseteq R_C$，这样得 $R_A=R_C\cup R_{[a]}\cup R_{A-C\cup[a]}=R_C\cup R_{A-C\cup[a]}=R_{A-[a]}$，所以 a 是类可约属性。

定理 3.9 设 (U,A,I) 是形式背景，则下列命题等价：

(1)$a\in A$ 是相对必要属性；

(2)$R_{\{a\}}\not\subset R_C$，$R_{\{a\}}\subseteq R_{A-\{a\}}$。

证明：由定理 3.6 和定理 3.7 可得。

4 形式背景中属性约简方法

可以利用类似粗糙集理论中辨识矩阵的方法来获得属性约简，只不过在这里用的是相似矩阵。

定义 4.1 设 (U,A,I) 是形式背景，对于 $x,y\in U$，记 $sim(x,y)=f_A(x)\cap f_B(x)$，称 $sim(x,y)$ 为 x 与 y 的相似属性集。称 $Sim=(sim(x,y):x,y\in U)$ 为形式背景 (U,A,I) 的相似属性矩阵。

定理 4.1 设 (U,A,I) 是形式背景，$B\subseteq A$，则 B 是属性协调集的充要条件是对于任意的相似属性集 $sim(x,y)\neq\phi$，都有 $B\cap sim(x,y)\neq\phi$。

证明：若 B 是属性协调集，即 $R_A\subseteq R_B$。对于任意的 $sim(x,y)\neq\phi$，有 $(x,y)\in R_A$，因此 $(x,y)\in R_B$，所以 $B\cap sim(x,y)=f_B(x)\cap f_B(y)\neq\phi$。反过来，若对于任意的相似属性集 $sim(x,y)\neq\phi$，都有 $B\cap sim(x,y)\neq\phi$，$\forall(x,y)\in R_A$，有 $sim(x,y)\neq\phi$，所以 $f_B(x)\cap f_B(y)=B\cap sim(x,y)\neq\phi$，即 $(x,y)\in R_B$，因此 $R_A\subseteq R_B$，得 B 是属性协调集。

定理 4.2 a 是核心属性的充要条件是存在 $x,y\in U$，满足 $sim(x,y)=\{a\}$。

证明：设 a 为核心属性，假设对任意 $sim(x,y)\neq\phi$，都有 $a\notin sin(x,y)$。则令 $B=\bigcup_{sim(x,y)\neq\phi}sim(x,y)$，易知 $a\notin B$，且由定理 4.1 可知 B 为属性协调集，所以必存在一个属性约简集 $D\subseteq B$，这时有 $a\notin D$，这与 a 核心属性相矛盾。又若含 a 的任何相似属性集都至少包含两个属性，则令 $B_1=\bigcup_{x,y\in U}(sim(x,y)-\{a\}$，则对于任意的 $sim(x,y)\neq\phi$，有 $B_1\cap sim(x,y)\neq\phi$。从定理 4.1 可知 B_1 为属性协调集，但是 $a\notin B_1$。由 B_1 是属性协调集，可知存在约简集 $D_i\subseteq B_1$，也有 $a\notin D_i$，这与 a 是核心属性相矛盾。综上，一定存在 $x,y\in U$，满足 $sim(x,y)=\{a\}$。反过来，设 $sim(x_{i_0},y_{j_0})=\{a\}$，若 $|A|=1$，则显然 a 是核心属性；若 $|A|>1$，则 $f_{A-\{a\}}(x_{i_0})\cap f_{A-\{a\}}(y_{j_0})=\phi$，即 $(x_{i_0},y_{j_0})\notin R_{A-\{a\}}$，但是 $(x_{i_0},y_{j_0})\in R_A$，因此 $R_A\not\subset R_{A-\{a\}}$，由定理 3.6 可知 a 是核心属性。

设 (U,A,I) 是形式背景，令 $\Delta=\wedge\{\vee(a_i:a_i\in sim(x,y)):x,y\in U,sim(x,y)\neq\phi\}$。根据定理 4.1 可知，$\Delta$ 的极小析取范式中每一合取项对应形式背景的一个属性约简集。因此，从 Δ 的极小析取范式可以得到所有的约简集。

例 1 在形式背景 (U,A,I) 中，$U=\{x_1,x_2,x_3,x_4,x_5\}$，$A=\{a_1,a_2,a_3,a_4,a_5\}$，$I$ 由表 1 给出。

表 1

	a_1	a_2	a_3	a_4	a_5
x_1	1	0	1	0	1
x_2	0	1	0	0	0
x_3	0	1	0	1	0
x_4	0	0	0	1	0
x_5	1	0	1	0	0

$sim(x_1,x_1)=\{a_1,a_3,a_5\}$，$sim(x_3,x_4)=sim(x_4,x_4)=\{a_4\}$，$sim(x_3,x_3)=\{a_2,a_4\}$，$sim(x_2,x_3)=sim(x_2,x_2)=\{a_2\}$，$sim(x_1,x_5)=sim(x_5,x_5)=\{a_1,a_3\}$，其余的相似属性集为空集。$\Delta=(a_1\vee a_3\vee a_5)\wedge(a_1\vee a_3)\wedge(a_2\vee a_4)\wedge a_2\wedge a_4$，化简得 $\Delta=(a_1\wedge a_2\wedge a_4)\vee(a_2\wedge a_3\wedge a_4)$。所以，形式背景 (U,A,I) 共有两个属性约简集 $\{a_1,a_2,a_4\}$ 和 $\{a_2,a_3,a_4\}$，核心属性集 $C=\{a_2,a_4\}$，相对必要属性集 $K=\{a_1,a_3\}$，不必要属性集 $N=\{a_5\}$。

结论 粗糙集理论是一种数据分析和知识发现的工具，已被成功应用于许多领域。本文在形式背景中定义了一种上、下近似的概念，给出了相应的属性约简的理论，即寻找保证论域中每一集合上、下近似不变的最小属性集。由于许多信息系统能够转化成形式背景，因此本文的结论可应用于一般的信息系统。本文工作对信息系统的知识发现有重要意义。文中给出了属性协调和属性约简的判定定理，引入了形式背景相似矩阵的概念，给出了基于相似矩阵的属性约简的判定方法，同时刻画了不同类型属性的特征。文中形式背景没有目标属性，对具有目标属性的形式背景可进一步进行研究。

参考文献

1 Pawlak Z. Rough set. International Journal of Computer and Information Science, 1982, 11(5): 342～356

2 Yao Y Y. Relational interpretations of neighborhood operators and rough set approximation. Information Sciences, 1998, 111: 239～259

3 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法. 北京：科学出版社，2001

4 张文修. Rough 集约简一般理论[J]. 见：2004 年全国第四届 Rough 集与软计算研讨会. 舟山，2004

5 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001

6 苗夺谦，胡贵容. 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

7 Li Tongjun, Zhang Wenxiu. Rough approximations in formal contexts. In: Proc. of Int Conf. on Machine Learning and Cybernetics, Guangzhou, China, 2005