决策表中基于条件信息熵的近似约简

杨 明
(南京师范大学计算机科学系,江苏南京 210097)

摘 要: 属性约简是粗糙集理论的重要研究内容,已有效应用于机器学习、数据挖掘等领域.基于条件信息熵的属性约简可有效推广代数观下的属性约简,但存在抗噪声弱且某些情况下冗余属性多的不足.为此,本文在引入决策表中基于条件信息熵的近似约简概念后,提出决策表中基于条件信息熵的近似约简算法,该算法可有效增强抗噪性,且可依据实际应用的需要有效地对冗余属性进行取舍.最后,本文侧重通过选择不同精度下的约简属性子集在 Benchmark 上进行了分类器的性能测试.

关键词: 粗糙集;属性约简;条件信息熵;近似约简
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2007) 11-2156-05

Approximate Reduction Based on Conditional Information Entropy in Decision Tables

YANG Ming
(*Department of Computer Science, Nanjing Normal University, Nanjing, Jiangsu* 210097, *China*)

Abstract: Attribute reduction is not only one of important parts researched in rough set theory, but also widely applied to many fields such as machine learning, data mining and so on. The attribute reduction method based on conditional information entropy can also be used effectively in the algebra view. However, these are two main disadvantages: this method is sensitive to noise and in some cases the obtained attribute subset may contain some redundant attributes. Therefore, in this paper, after introducing a concept of approximate reduction based on conditional information entropy in decision tables, we present an approximate reduction algorithm based on conditional information entropy (ARABCIE). The algorithm can effectively improve sensitivity to noise and properly select those redundant attributes by applications. Finally, we discuss the robustness of ARABCIE algorithm by experimenting on benchmark using several attribute subsets with different precision.

Key words: rough set; attributes reduction; conditional information entropy; approximate reduction

1 引言

波兰数学家 Z. Pawlak 80 年代初提出的 Rough Set (RS,粗糙集)是一种新的处理不精确、不完全与不相容知识的数学理论[1],近年来该理论在机器学习、数据挖掘及模式识别等多个领域得到了广泛的应用[2,3].属性约简是粗糙集的重要内容,已有效地应用于机器学习[4,5]、数据挖掘[6]等多个领域,如可用于机器学习中的特征选择及数据挖掘中的数据预处理.

基于粗糙集方法的属性约简就是发现一个特征子集(特征子集就是一个属性约简,这里特征即为属性),由于属性约简删除了信息系统中的冗余属性,仅保留其重要属性,因而和数据挖掘算法相结合可有效减少挖掘的规则数[6].可见,粗糙集理论中的属性约简算法研究具有十分重要的理论和应用意义.

研究者主要从代数观出发对粗糙集理论进行研究,在属性核求解和属性约简算法研究上取得了一定的进展[7~9,14],但从信息观出发进行属性约简研究相对少一些.王国胤教授对信息观下属性约简进行深入的研究[10,11],在和代数观下属性约简进行深入比较分析后,得到两种观点下的一些重要等价性质和主要的不同特性,进而提出基于条件信息熵的决策表属性约简算法.不同于代数观下的属性约简,一致对象划分和不一致对象划分均可改变条件信息熵.对不一致决策表而言,基于条件信息熵的属性约简可有效地增加一些影响不一致对象划分粒度的属性,因而不仅为不一致决策表的知识获取提供更加有效的途径,而且也从信息观角度揭示了属性约简的本质.因此,对条件信息熵下的属性约简

收稿日期:2006-12-18;修回日期:2007-06-18
基金项目:国家自然科学基金(No.40771163);江苏省自然科学基金(No.BK2005135);江苏省高校自然科学研究项目基金(No.05KJB520066)

进行进一步的研究具有十分重要的意义.

在深入分析和研究条件信息熵的属性约简后,可发现该方法存在以下不足:抗噪声干扰能力弱;某些情况下会包含过多的冗余属性,这里的冗余属性主要指:为进一步划分不一致对象需要增加的部分属性,也包括在得到的特征子集中那些对分类贡献相对较小的属性.为此,本文在引入决策表中基于条件信息熵的近似约简概念后,提出决策表中基于条件信息熵的近似约简算法,该算法可依据实际应用的需要有效地对冗余属性进行取舍,且可有效增强抗噪声干扰的能力,因而是基于条件信息熵的决策表约简的推广和补充.实验结果表明本文算法可通过调节精度得到有效改进分类性能的约简属性子集.

2　粗糙集的信息论观点和相关结论

为节省篇幅,本文主要介绍在信息观下属性约简的一些概念[10,11],关于粗糙集理论在代数观下的一些概念可参见文献[2,3].

决策表 DT 是一个四元组$\langle U, C\cup D, V, f\rangle$,其中,$U$ 是一组对象的非空有限集合,称为论域;设有 n 个对象,则 U 可表示为:$U=\{x_1, x_2, \cdots, x_n\}$,$C$ 为条件属性集,D 为决策属性集,通常 D 仅包含一个决策属性.$V=\bigcup\limits_{a\in C\cup D} V_a$,$V_a$ 为属性 a 的值域集;f 是 $U\times(C\cup D)\to V$ 的映射.对 $B\subseteq(C\cup D)$,无差别关系 $IND(B)$ 定义为 $\{(x,y)\in U^2 \mid \forall a\in B, f(x,a)=f(y,a)\}$,通过 $IND(B)$ 将 U 划分为若干个类 $E_i(1\leqslant i\leqslant |U/IND(B)|)$.为便于叙述,用 $|\cdot|$ 表示集合的基.

设 X 为论域 U 的一个子集,$P\subseteq C$,X 的关于 P 的下近似为 $\underline{P}X=\{x\in U:[x]_P\subseteq X\}$,其中,$[x]_P$ 表示 U 中所有与 x 在关系 $IND(P)$ 下是等价的元素构成的集合.

定义 1　设 P,Q 在 U 上导出的划分分别为 $X, Y(X=\{X_1, X_2, \cdots, X_s\}, Y=\{Y_1, Y_2, \cdots, Y_t\})$,则 P,Q 在 U 的子集组成的 σ 代数上的概率分布为

$$[X:P]=\begin{bmatrix} X_1 & X_2 & \cdots & X_s \\ p(X_1) & p(X_2) & \cdots & p(X_s)\end{bmatrix}$$

$$[Y:P]=\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_t \\ p(Y_1) & p(Y_2) & \cdots & p(Y_t)\end{bmatrix}$$

其中,$p(X_i)=\dfrac{|X_i|}{|U|}$,$i=1,2,\cdots,s$;$p(Y_j)=\dfrac{|Y_j|}{|U|}$,$j=1,2,\cdots,t$.

定义 2　知识(属性集合)$Q(U\mid IND(Q)=\{Y_1, Y_2, \cdots, Y_t\})$相对于知识(属性集合)$P(U\mid IND(P)=\{X_1, X_2, \cdots, X_s\})$的条件熵 $H(Q\mid P)$ 定义为

$$H(Q\mid P)=-\sum_{i=1}^{s} p(X_i)\sum_{j=1}^{t} p(Y_j\mid X_i)\log(p(Y_j\mid X_i)),$$

其中,$p(Y_j\mid X_i)=|Y_j\cap X_i|/|X_i|$,$i=1,2,\cdots,s$;$j=1,2,\cdots,t$.

定义 3　(条件信息熵下的属性约简定义)　对给定的决策表 DT,若 $H(D\mid C)=H(D\mid A)(A\subseteq C)$,且 $H(D\mid C)\neq H(D\mid B)(\forall B\subset A)$,则 A 为决策表的一个属性约简.

定义 4　设 $P\subseteq C$,对 D 的划分 $\{\Psi_1, \Psi_2, \cdots, \Psi_k\}$ 的 P-近似精度为 $\gamma_P=|POS_P(D)|/|U|$,其中 $POS_P(D)=\bigcup\limits_{i=1}^{k}\underline{P}\Psi_i$.

定义 5　(代数观下的属性约简定义)设 $P\subseteq C$,若 $\gamma_P=\gamma_C$,则称 P 为 C 的一个(相对于决策属性 D 的)候选属性约简.若 P 为候选属性约简,且不存在 $R\subset P$,使得 $\gamma_R=\gamma_C$,则称 P 为属性约简.

文献[10,11]对代数观和信息观两种观点下的属性约简进行了比较分析,证明了在一致决策表情况下定义 3 和定义 5 是等价的;而在不一致决策表情况下定义 3 和定义 5 不等价.由文献[10]的结论可得如下的引理 1 和引理 2.

引理 1　若 $H(D\mid A\cup\{a\})=H(D\mid A)$,则 $POS_{A\cup\{a\}}(D)=POS_A(D)$.

引理 2　若 $A\subseteq C$,$B\subseteq C$,且 $A\subseteq B$,则 $H(D\mid A)\geqslant H(D\mid B)$.

定理 1　若 A 是条件信息熵下的属性约简,则 A 是代数观下的候选属性约简,但未必是代数观下的属性约简.

证明:因 A 是条件信息熵下的属性约简,即 $H(D\mid C)=H(D\mid A)(A\subseteq C)$,且 $H(D\mid C)\neq H(D\mid B)(\forall B\subset A)$,故由引理 1 知 $POS_C(D)=POS_A(D)$,即 A 是代数观下的候选属性约简,但未必是代数观下的属性约简(见实例 1 的说明).证毕.

由定理 1 可知,对任意给定的条件信息熵下的属性约简 A,一定存在一个代数观下的属性约简 $B\subseteq A$,显然有 $|A|\geqslant|B|$.也就是说,条件信息熵下的属性约简相对较长一些,可能会包含一些冗余的属性,这些属性加入的主要目的是进一步划分不一致对象,从而改变不一致对象的划分,使得条件信息熵减小.这些可有效改变不一致对象划分的属性加入增强不一致对象之间鉴别,即该对象隶属于哪个类别的可能性更大.为此,我们通过下面的实例 1 来加以说明.

对实例 1,由定义 2 可得:$H(\{d\}\mid\{C1, C2, C3\})=0.52707$,$H(\{d\}\mid\{C1\})=0.611829$,$H(\{d\}\mid\{C2\})=0.605536$,$H(\{d\}\mid\{C1, C2\})=0.53337$,$H(\{d\}\mid\{C1, C3\})=0.554312$.因此,依据文献[10]基于条件信息熵的属性约简算法可得信息观下的属性约简为 $\{C1, C2, C3\}$;而按照代数观下的属性约简算法可得代数观下的属性约简为 $\{C1\}$,且 $\{C1\}$ 也是代数观下的属性核.可

见,对不一致决策表,条件信息熵下的属性约简中包含一个代数观的属性约简.

实例 1 表 1 所示的是一决策表,其中共有 11 个元素和 4 个属性,$C=\{C1,C2,C3\}$ 为条件属性集,d 为决策属性.

表 1 决策表

元素 \ 属性	$C1$	$C2$	$C3$	d
$x1$	1	0	0	0
$x2$	0	0	0	0
$x3$	0	0	0	1
$x4$	0	1	0	0
$x5$	0	1	0	1
$x6$	0	1	0	0
$x7$	0	1	0	0
$x8$	0	1	0	0
$x9$	0	0	1	0
$x10$	0	0	1	1
$x11$	0	0	1	1

从实例 1 可以看出,$\{C1\}$ 可有效将确定对象区分开,而将不一致对象划分为一个等价类.条件信息熵下的属性约简算法以 $\{C1\}$ 为出发点,逐次增加 $\{C2\}$ 和 $\{C3\}$ 以便进一步划分不一致对象,减小条件信息熵.可见,条件信息熵下的属性约简可有效增强不一致对象区分,但是以增加更多的属性为代价.因此,如何使条件信息熵尽可能小,且有效地控制冗余属性的增加是本文的主要目的.

此外,由下面的定理 2 可知,对一个含 n 个样本的决策表,若有一个含有 p 个样本的等价类中含有 $m-1$ 类噪声,它们的数目分别是 $x_1,\cdots,x_{m-1}$,且 $\sum_{i=1}^{m-1} x_i \ll x_m$,即 $m-1$ 类噪声的样本总数很小,则当 n 规模较大或 p 规模相对较大时,该等价类对应的条件信息熵很小,但不为 0.可见,无论是对一致决策表还是对不一致决策表而言,若含有噪声,则定义 3 中的"$H(D|C)=H(D|A)(A\subseteq C)$"要求显得过于严格.于是,在本文的第 3 部分引入条件信息熵下的 ε-近似属性约简定义(详见第 3 部分).

定理 2

若 $f(x_1,x_1,\cdots,x_m)=-\frac{p}{n}\sum_{i=1}^{m}\frac{x_i}{p}\log_2\left(\frac{x_i}{p}\right)$,其中,$n>p>0$,$\sum_{i=1}^{m} x_i=p$ 且 $\sum_{i=1}^{m-1} x_i \ll x_m$,

则当 $n\to\infty$ 或 $p\to\infty$ 时有 $f(x_1,x_2,\cdots,x_m)\to 0$.

证明:由已知条件知 $f(x_1,x_2,\cdots,x_m)=\frac{p}{n}(\log_2(p)\cdot\sum_{i=1}^{m} x_i-\frac{1}{p}(\sum_{i=1}^{m-1} x_i\log_2(x_i)+x_m\log_2(x_m)))$,而 $\sum_{i=1}^{m} x_i=p$,故 $f(x_1,x_2,\cdots,x_m)=\frac{p}{n}(\log_2(p)-\frac{1}{p}x_m\log_2(x_m)-\frac{1}{p}\cdot\sum_{i=1}^{m-1} x_i\log_2(x_i))$;由 $x_i\geqslant 1,i=1,2,\cdots,m$,$\sum_{i=1}^{m-1} x_i\ll x_m$ 知 $\sum_{i=1}^{m-1} x_i\log_2(x_i)\ll x_m\log_2(x_m)$,$f(x_1,\cdots,x_m)\approx\frac{p}{n}(\log_2(p)-\frac{x_m}{p}\log_2(x_m))$.因此,当 $n\to\infty$ 或 $p\to\infty$ 时有 $f(x_1,x_2,\cdots,x_m)\approx\frac{p}{n}(\log_2(p)-\frac{x_m}{p}\log_2(x_m))\to 0$.证毕.

3 条件信息熵下的决策表近似约简

为增强抗噪声的能力,且有效地控制冗余属性的增加且使条件信息熵尽可能小(即满足一定的精度),本文引入下面的定义 6.

定义 6 (条件信息熵下的 ε-近似属性约简定义)对给定的决策表 DT,对给定 $\varepsilon(\varepsilon\geqslant 0)$,若 $|H(D|C)-H(D|A)|\leqslant\varepsilon(A\subseteq C)$,且 $|H(D|C)-H(D|B)|>\varepsilon(\forall B\subset A)$,则 A 为决策表的一个 ε-近似属性约简.

定理 3 对给定的决策表 DT,条件信息熵下的 ε-近似属性约简是条件信息熵下的属性约简的推广.

证明:由定义 6 知,当 $\varepsilon=0$ 时,定义 6 和定义 3 等价,因而定义 6 是文献[10,11]的推广.证毕.

对实例 1,当取 $\varepsilon=0.01$,则由 $|H(\{d\}|\{C1,C2\})-H(\{d\}|\{C1,C2,C3\})|=0.0073<\varepsilon$,而 $|H(\{d\}|\{C1\})-H(\{d\}|\{C1,C2,C3\})|=0.084752>\varepsilon$,$|H(\{d\}|\{C2\})-H(\{d\}|\{C1,C2,C3\})|=0.078459>\varepsilon$;依据定义 6,$\{C1,C2\}$ 为条件信息熵下的一个 ε-近似属性约简.可见,通过参数的有效设置可有效地进行属性的合理取舍.

类似文献[10],依据条件信息熵下重要属性的度量准则,逐步选择重要属性,直到发现一个满足用户给定精度要求的属性约简,具体算法步骤如下.

算法 1 ARABCIE(Approximate Reduction Algorithm Based on Conditional Information Entropy)//基于条件熵的近似约简算法

输入:一个决策表 $DT=\langle U,C\cup D,V,f\rangle$,其中,$U$ 为论域,C,D 分别为条件属性集和决策属性集;$\varepsilon(\varepsilon\geqslant 0)$为给定的可调参数;

输出:该决策表的一个近似属性约简 B.

步骤 1 计算条件属性集 C 中决策属性集 D 相对条件属性集 C 的条件熵 $H(D|C)$.

步骤 2 计算条件属性集 C 中相对决策属性集 D 的核属性集 $Core$,并令 $Att=C-Core$.

步骤 3 令 $B=Core$,

步骤 3.1 如果 $|B|>0$,则计算条件熵 $H(D|B)$,转步骤 3.4;

步骤 3.2 对每个属性 $a_i\in Att$,计算决策属性集 D 相对条件属性集 $B\cup\{a_i\}$ 的条件熵 $H(D|B\cup\{a_i\})$;

步骤 3.3 选择使 $H(D|B\cup\{a_i\})$ 最小的属性 a_j(若同时有多个属性达到最小值,则从中选择一个与 B 的属性值组合数最少的属性),$Att=Att-\{a_j\}$,$B=B\cup\{a_j\}$;

步骤 3.4 若 $|H(D|B)-H(D|C)|\leqslant\varepsilon$,则转步骤 4,否则转步骤 3.2.

步骤 4　按 $H(D|\{a_i\})$($a_i \in B$)递减的顺序对每个 a_i 重复下述操作:

步骤 4.1　计算决策属性集相对条件属性集 B 在删掉 a_i 后的条件熵 $H(D|\{a_i\})$;

步骤 4.2　如果 $|H(D|B-\{a_i\})-H(D|C)|\leqslant\varepsilon$,则属性 a_i 从 B 中删除,$B=B-\{a_i\}$;否则,属性 a_i 不能被约简,B 不变.

由算法 1 可发现满足定义 6 要求的近似属性约简,且算法 1 的时间复杂度与文献[10,11]相应属性约简算法的时间复杂度相当.

当然,如何选择有效的 ε 值需要依据实际应用的需要进行确定.在实际应用中,ε 值的大小可依据原有属性个数和总的条件信息熵进行合理的确定,因为属性选择过多或过少均会影响分类器性能,过多会影响分类器的推广性,过少会影响分类器的精度.为此,下面的第 4 部分将通过实验来验证不同的 ε 值对属性约简及分类精度的影响.

4　实验结果

为进一步验证算法的有效性,对本文的 ARABCIE 算法和文献[10,11]中的 CEBARKCC 算法用 VC ++ 进行了实现,并选用 UCI 机器学习数据库中的 6 个数据集进行了实验测试.选择的 6 个数据集描述如下:(1)Contraceptive Method Choice 数据集,共有 1473 个样本,移去第 9 个属性并将 2、3 类合并为 1 类,形成具有 8 个条件属性和 1 个决策属性的数据集,简记为 CMC;(2)pima Indians diabetes 数据集,共有 768 个样本、8 个条件属性和 1 个决策属性(共有 2 类),简记为 Diabetes;(3)Glass Identification 数据集,共有 214 个样本,将其分成 float 和 non float 类(即得 2 类),移去其第 1 列(即 ID number)后得到的数据集作为实验数据,简记为 Glass;(4)Ionosphere 数据集,共有 351 个样本、34 个条件属性和 1 个决策属性(共 2 类),记为 Ionosphere;(5)iris 数据集,共有 150 个样本、4 个条件属性和 1 个决策属性,含有 3 类,每类 50 个样本.取第 2、3 类线性不可分数据作为实验数据,简记为 Iris23;(6)BUPA liver disorders 数据集,共有 345 个样本、6 个条件属性和 1 个决策属性(共 2 类),简记为Liver.

为方便计,将各数据集的条件属性按其列号依次记为 $a_1,a_2,\cdots$,并将 ε-近似属性约简 B 记为(B,ε).在实验中,随机选取各数据集的 80% 用于属性约简,余下的 20% 作为测试;同时,为有效地估计约简属性子集的分类性能,我们采用 S.Mika 提出的 KFDA 算法对约简得到的各属性子集进行分类器设计,这里采用 RBF 核,实验结果如表 2 所示.

从表 2 可以看出,使用文献[10]的 CEBARKCC 算法在多个数据集上,可有效地进行属性约简且保持分类能力不变或是可比较的.同时,我们也发现在某些情况下 CEBARKCC 算法得到的约简未必是最有效的,而在文献[10,11]的基础上,提出的改进算法 ARABCIE 通过可调精度参数 ε 选择可得到更有效的约简属性;如:对 CMC 数据集,选择 ε=0.05 得到的属性约简提高了分类能力,且比 CEBARKCC 算法得到的约简属性子集规模小;对 Glass 数据集,选择 ε=0.05 得到的属性约简进一步缩小了约简属性集的规模,且识别率明显提高,由 95.45% 提高到 97.27%;在 Iris23 数据集上,ARABCIE 算法保持识别率相同的情况下,进一步降低了约简属性集的规模.

表 2　属性约简测试结果

数据集	CEBARKCC 算法	ARABCIE 算法	KFDA 算法	
	属性约简	属性约简	属性集	识别率
CMC	$\{a_1,a_2,a_3,a_6,a_7,a_8\}$		all	92.54%
		$(\{a_1,a_2,a_3,a_6,a_7,a_8\},0.005)$	$\{a_1,a_2,a_3,a_6,a_7,a_8\}$	91.86%
		$(\{a_1,a_2,a_6,a_7,a_8\},0.01)$	$\{a_1,a_2,a_6,a_7,a_8\}$	92.20%
		$(\{a_1,a_2,a_7,a_8\},0.05)$	$\{\mathbf{a_1,a_2,a_7,a_8}\}$	**92.20%**
Diabetes	All	$(all,0.001)$	all	71.42%
		$(\{a_1,a_2,a_3,a_5,a_6,a_7,a_8\},0.01)$	$\{\mathbf{a_1,a_2,a_3,a_5,a_6,a_7,a_8}\}$	**71.42%**
		$(\{a_1,a_2,a_3,a_5,a_6,a_7\},0.03)$	$\{a_1,a_2,a_3,a_5,a_6,a_7\}$	70.13%
		$(\{a_2,a_3,a_5,a_6,a_7\},0.05)$	$\{a_2,a_3,a_5,a_6,a_7\}$	70.13%
Glass	$\{a_1,a_3,a_5,a_6,a_7\}$		all	95.45%
		$(\{a_1,a_3,a_5,a_6,a_7\},0.001)$	$\{a_1,a_3,a_5,a_6,a_7\}$	95.45%
		$(\{a_1,a_3,a_6,a_7\},0.05)$	$\{\mathbf{a_1,a_3,a_6,a_7}\}$	**97.27%**
		$(\{a_1,a_3,a_6\},0.1)$	$\{a_1,a_3,a_6\}$	95.45%
Ionosphere	$\{a_{14},a_{16},a_{28}\}$		all	88.73%
		$(\{a_{14},a_{16},a_{28}\},0.005)$	$\{\mathbf{a_{14},a_{16},a_{28}}\}$	**87.33%**
		$(\{a_{16},a_{28}\},0.01)$	$\{a_{16},a_{28}\}$	84.50%
Iris23	$\{a_1,a_3,a_4\}$		all	100%
		$(\{a_1,a_3,a_4\},0.01)$	$\{a_1,a_3,a_4\}$	100%
		$(\{a_3,a_4\},0.05)$	$\{\mathbf{a_3,a_4}\}$	**100%**
Liver	all	$(a_l1,0.001)$	all	63.77%
		$(\{a_1,a_2,a_3,a_4,a_5\},0.005)$	$\{\mathbf{a_1,a_2,a_3,a_4,a_5}\}$	**63.77%**
		$(\{a_2,a_3,a_4,a_5\},0.02)$	$\{a_2,a_3,a_4,a_5\}$	62.32

注:其中,all 表示所有的条件属性集;黑体标注的约简为具有高设别率且规模较小的约简.

可见,通过精度参数 ε 的选择,ARABCIE 算法可有效降低约简属性集的规模,改进分类器的性能,因此是算法 CEBARKCC 算法的补充和改进.当然,在实际应用中,如何更加有效选择精度参数 ε 将是我们的未来研究内容之一.

5　结语

本文在引入决策表中基于条件

信息熵的近似约简概念后,提出决策表中基于条件信息熵的近似约简算法,该算法可有效过滤噪声,且可依据实际应用的需要有效地对冗余属性进行取舍,因而有效地推广了基于条件信息熵的属性约简方法.

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science, 1982, 11(5): 341 – 356.

[2] Pawlak Z. Rough set approach to multi-attribute decision analysis [J]. European Journal of Operational Research, 1994, 72(3): 443 – 459.

[3] 刘清. Rough 集及 Rough 推理[M]. 北京:科学出版社, 2001.
Liu Qing. Rough Sets and Rough Reasoning[M]. Beijing: Science Press, 2001. (in Chinese)

[4] Swiniarski R W, Skowron A. Rough set methods in feature selection and recognition[J]. Pattern Recognition Letters, 2003, 24 (6): 833 – 849.

[5] Jensen R, Shen Q. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1457 – 1471.

[6] Li J Y, Cercone N. A rough set based model to rank the importance of association rules[A]. Proceedings of 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Lecture Notes in Computer Science[C]. Regina, Canada: Springer Berlin/Heidelberg, 2005. 109 – 118.

[7] 杨明. 一种基于改进差别矩阵的核增量式更新算法[J]. 计算机学报, 2006, 29(3): 407 – 413.
Yang Ming. An incremental updating algorithm of the computation of a core based on the improved discernibility matrix[J]. Chinese Journal of Computers, 2006, 29(3): 407 – 413. (in Chinese)

[8] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报, 2007, 30(5): 815 – 822.
Yang Ming. An incremental updating algorithm for attribute reduction based on the improved discernibility matrix[J]. Chinese Journal of Computers, 2007, 30(5): 815 – 822. (in Chinese)

[9] Wang Jue, Wang Ju. Reduction algorithm based on discernibility matrix the ordered attributes method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489 – 504.

[10] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759 – 766.
Wang Guoyin, Yu Hong, Yang Dachun. Decision table reduction on conditional information entropy[J]. Chinese Journal of Computers, 2002, 25(7): 759 – 766. (in Chinese)

[11] Wang Guo Yin, et al. Theoretical study on attribute reduction of rough set theory: comparison of algebra and information views [A]. Proceedings of the Third IEEE International Conference on Cognitive Informatics[C]. Canada: IEEE Computer Society, 2004. 148 – 155.

[12] 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 Duo-Qian, Hu Gui-Rong. A heuristic algorithm for reduction of knowledge[J]. Journal of Computer Research & Development, 1999, 36(6): 681 – 684. (in Chinese)

[13] Mika S, Ratsch G, Weston J, B, et al. Fisher discriminant analysis with kernels[A]. Neural Networks for Signal Processing IX, Proceedings of the 1999 IEEE Signal Processing Society Workshop[C]. Y-H Hu, J Larsen, E Wilson, S Douglas, Eds, New York: IEEE Press, 1999. 41 – 48.

[14] 叶东毅. Jelonek 属性约简算法的一个改进[J]. 电子学报, 2000, 28(12): 81 – 82.
Ye Dong-yi. An improvement to Jelonek's attribute reduction algorithm[J]. Acta Electronica Sinica, 2000, 28(12): 81 – 82. (in Chinese)

作者简介:

杨 明 男, 1964 年 11 月生, 博士, 教授, 东南大学计算机应用专业博士毕业. 主要研究方向为数据挖掘和知识发现, 机器学习, 粗集理论与应用. E-mail: m.yang@njnu.edu.cn