基于核属性的知识获取算法*)

朱振国[1] 赵 毅[2] 李 邕[3]

(重庆交通大学计算机及信息学院 重庆 400074)[1] (重庆交通大学教育技术中心 重庆 400074)[2]

(重庆邮电大学计算机科学与技术研究所 重庆 400065)[3]

摘 要 Rough集理论是近年来发展起来的一种有效地处理不精确、不确定、含糊信息的数学理论方法,在机器学习、数据挖掘、智能数据分析、控制算法获取等领域取得了很大的成功。决策表是Rough Set理论的处理对象,用Rough Set对决策表进行规则提取通常有代数观和信息观两种主要理论和方法,使用哪一种方法提取的规则集更好是很多研究者的目标。本文针对Rough Set理论的核心内容之一的知识获取进行了研究,提出了一种基于属性重要性排序的知识获取算法,并且证明了在不相容系统中使用信息观方法比使用代数观的方法更好,能够提取更合理的规则集。

关键词 Rough Set,核属性,代数观,信息观

Knowledge Acquisition Algorithm Based on Attribute Core

ZHU Zhen-guo[1] ZHAO Yi[2] LI Yong[3]

(School of Computer Science and Information, Chongqing Jiaotong University, Chongqing 400074, China)[1]

(Educational Technology Center, Chongqing Jiaotong University, Chongqing 400074, China)[2]

(Institute of Computer Science and Technology, Chongqing University of Posts & Telecommunications, Chongqing 400065, China)[3]

Abstract Rough set is a valid mathematical theory developed in recent years, which has the ability to deal with imprecise, uncertain, and vague information. It has been applied in such fields as machine learning, data mining, intelligent data analyzing and control algorithm acquiring successfully. The attribute core of a decision table is often the start point and key of many decision information system reduction procedures based on rough set theory. The algebra view and information view are two main views and of rough set theory in knowledge acquisition. Many scholars are finding the best methods which can get better rules using algebra view or information view. In this paper, based on the problem of rules acquisition of a decision table, we develop a knowledge acquisition algorithm based on attribute sorting according to importance of attribute, and prove that we can get better rules using information view in incompatible system.

Keywords Rough set, Attribute core, Algebra view, Information view

1 引言

Rough Set理论是波兰科学家Z. Pawlak于1982年提出的一种处理不精确、不相容和不完全数据的新的数学工具。知识约简是Rough Set理论研究的核心内容之一。一般来讲,一个决策表的知识相对约简不是唯一的,人们期望找到具有最少属性的约简,即最小约简。然而,遗憾的是Wong. S. K. M和Ziarko. W已经证明找出一个决策表的最小约简是NP-hard问题[1]。

决策表核属性的确定往往是信息约简的基础。在Rough Set理论的研究中有两种主要的理论观点和方法,即Rough Set理论的代数观(传统的Pawlak对于Rough Set理论的定义)和信息观(Rough Set理论的信息熵定义)[3,4]。本文结合前人对Rough Set理论的代数观和信息观的研究成果[5-13],提出了一种基于属性重要性排序的知识获取算法,并且证明了在不相容系统中使用信息观方法能够提取好的规则集。

2 代数观下的核属性计算[1]

Hu X H根据Skowron教授提出的可辨识矩阵得出了一个确定决策表信息系统核属性集合的方法[8]。

定义1 一个决策表信息系统 $S=\langle U,R,V,f\rangle$。其中:U为论域,$R=C\cup D$,C和D分别为条件属性集和结果属性集,$D\neq\Phi$;$V=\bigcup_{r\in R}V_r$为属性值的集合,V_r表示属性$r\in R$的值域;$f:U\times R\to V$是一个信息函数,它指定U中每一个对象x的属性值。

定义2 给定决策表信息系统$S=\langle U,R,V,f\rangle$,对于每个子集$X\subseteq U$和不分明关系B,X的下近似集和上近似集可以分别定义为

$$B-(X)=\bigcup\{Y_i\mid(Y_i\in U\mid IND(B)\wedge Y_i\subseteq X)\}$$

$$B-(X)=\bigcup\{Y_i\mid(Y_i\in U\mid IND(B)\wedge Y_i\cap X\neq\phi)\}$$

其中,$U\mid IND(B)=\{X\mid(X\subseteq U\underset{x\in X,y\in X,b\in B}{\forall}(b(x)=b(y)))\}$是不分明关系$B$对$U$的划分。

定义3 设U为一个论域,P,Q为U上的两个等价关系簇,若P的Q独立子集$S(S\subset P)$有$POSS(Q)=POSP(Q)$,则称S为P的Q约简。

可以记P的所有Q约简关系簇为$REDQ(P)$。

定义4 设U为一个论域,P,Q为U上的两个等价关系簇,$REDQ(P)$为P的所有Q约简关系簇,$COREQ(P)$为P的Q核,则$COREQ(P)=\bigcap REDQ(P)$。

定义5 令决策表系统为$S=\langle U,R,V,f\rangle$,$R=C\cup D$是属性集合,子集$C=\{a_i\mid i=1,\cdots,m\}$和$D=\{d\}$分别称为条件属性集和决策属性集,$U=\{x_1,x_2,\cdots,x_n\}$是论域,$a_i(x_j)$是样

*)基金项目:重庆交通大学青年科学基金(200533)。朱振国 硕士,讲师,主要从事模式识别、数据挖掘、网络安全的研究;赵 毅 硕士,副研究员,主要从事网络安全、数据挖掘研究;李 邕 硕士,主要从事智能信息处理研究。

本 x_j 在属性 a_i 上的取值。$CD(i,j)$表示可辨识矩阵中第 i 行 j 列的元素，则可辨识矩阵 CD 定义为：

$$C_D(i,j)=\begin{cases}\{a_k \mid a_k\in P \wedge a_k(x_i)\neq a_k(x_j)\}, & d(x_i)\neq d(x_j)\\ 0, & d(x_i)=d(x_j)\end{cases}$$

其中，$i,j=1,\cdots,n$。

文献[8]得出了如下结论：当且仅当某个 $CD(i,j)$为单属性集合时，该属性属于核 $CORED(C)$。

文献[2]在 Skowron 教授提出的可辨识矩阵的基础上，提出了一种改进的可辨识矩阵，并通过该矩阵提出了一种计算核属性的方法。

定义 6 给定信息系统 S，可辨识矩阵 C'_D 的元素 $C'_D(i,j)$定义为：

$$C'_D(i,j)=\begin{cases}C_D(i,j) & ,\min\{|D(x_i)|,|D(x_j)|\}=1\\ \varphi & ,\text{else}\end{cases}$$

文献[2]的结论如下：当且仅当某个元素 $C'_D(i,j)$为单属性集合时，该属性属于决策表的核 $CORED(C)$。

可以证明，叶东毅的求核方法所得到的结果是决策表在代数观中的核属性集[5]。

3 信息观下的核属性计算[1]

定理 1 设 U 是一个论域，P 是 U 的一个条件属性集合，d 为决策属性，且论域 U 是在 P 上相对于$\{d\}$一致的，则 P 中的一个属性 r 是 P 相对于决策属性 d 不必要的(多余的)，其充分必要条件为

$H(\{d\}\mid P)=H(\{d\}\mid P-\{r\})$。

定理 2 设 U 是一个论域，P 是 U 的一个条件属性集合，d 为决策属性，且论域 U 是在 P 上相对于$\{d\}$一致的，则 P 是相对于决策属性 d 独立的，其充分必要条件为对于 P 中任意属性 r 都有 $H(\{d\}\mid P)\neq H(\{d\}\mid P-\{r\})$成立。

定理 3 设 U 是一个论域，P 是 U 的一个条件属性集合，d 为决策属性，且论域 U 是在 P 上相对于$\{d\}$一致的，则 $Q\subseteq P$ 是 P 相对于决策属性 d 的一个约简的充分必要条件为 $H(\{d\}\mid Q)=H(\{d\}\mid P)$；且 Q 是相对于决策属性 d 独立的。

上述 3 个定理说明，对于一致的决策表，可以根据条件熵建立信息熵观点和代数观点对于约简的统一描述。但是，对于不一致的决策表，情况就发生了变化。下面给出决策表约简和核属性在信息熵观点中的一般定义。

定理 4 设论域为 U，某个等价关系在 U 上形成的划分为 $A_1=\{X_1,X_2,\cdots,X_n\}$，而

$$A_2=\{X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_{j-1},X_{j+1},\cdots,X_n,X_i\cup X_j\}$$

是将划分 A_1 中的某两个等价块 X_i 与 X_j 合并为 $X_i\cup X_j$ 得到的新划分。$B=\{Y_1,Y_2,\cdots,Y_m\}$也是 U 上的一个划分，且记

$$H(B\mid A_1)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j\mid X_i)\log(p(Y_j\mid X_i)),$$

$$\begin{aligned}H(B\mid A_2)=H(B\mid A_1)&-p(X_i\cup X_j)\sum_{k=1}^{m}p(Y_k\mid X_i\cup X_j)\\ &\log(p(Y_k\mid X_i\cup X_j))+p(X_i)\sum_{k=1}^{m}p(Y_k\mid X_i)\\ &\log(p(Y_k\mid X_i))+p(X_j)\sum_{k=1}^{m}p(Y_k\mid X_j)\\ &\log(p(Y_k\mid X_j))\end{aligned}$$

则 $H(B\mid A_2)\geqslant(H(B\mid A_1)$。

定理 4 说明，如果将决策表条件属性集的分类进行合并，将导致条件熵的上升，只有在发生合并的两个分类对于决策类的隶属度(概率)均相等的情况下，才可能不导致条件熵的变化。

由定理 4 我们得到下列推论：

推论 1 设 $T=\langle U,C\cup D,V,f\rangle$为一个决策表，$C$ 为初始条件属性集，D 为决策属性集，$a_i\in C,i=1,2,3,\cdots,m(m=|C|)$，则有：$H(D\mid\{a_1\})\geqslant H(D\mid\{a_1\}\cup\{a_2\})\geqslant\cdots\geqslant H(D\mid\{a_1\}\cup\cdots\cup\{a_i\}\cup\cdots\cup\{a_m\})=H(D\mid C)$。

证明：因为划分 $U\mid IND(\{a_1\})$是可以通过将划分 $U\mid IND(\{a_1\}\cup\{a_2\})$中的部分等价块合并得到的，根据上面的定理 1 可知，$H(D\mid\{a_1\})\geqslant H(D\mid\{a_1\}\cup\{a_2\})$；同理：$H(D\mid\{a_1\})\geqslant H(D\mid\{a_1\}\cup\{a_2\})\geqslant H(D\mid\{a_1\}\cup\{a_2\}\cup\{a_3\})\geqslant\cdots\geqslant H(D\mid\{a_1\}\cup\cdots\cup\{a_i\}\cup\cdots\cup\{a_m\})=H(D\mid C)$。

推论 1 说明，在知识约简中决策属性集相对条件属性集的条件熵(以下简称条件熵)的变化规律呈现非严格单调性。

推论 2 设 $T=\langle U,C\cup D,V,f\rangle$为一个决策表，$C$ 为初始条件属性集，D 为决策属性集，B 为知识约简后得到的条件属性集，C_0 为决策表的核。如果 $a_i\in C-C_0$ 是任意一个不能被约简的属性，且 $B\supset C_0$，则有：$H(D\mid C_o)>H(D\mid C_o\cup\{a_1\})>\cdots>H(D\mid C_o\cup\{a_1\}\cup\cdots\cup\{a_i\}\cup\cdots)>\cdots>H(D\mid B)$。

证明：由于 a_1 是一个不能被约简的属性，由定理 4 知：$H(D\mid C_o)>H(D\mid C_o\cup\{a_1\})$；同理 $H(D\mid C_o\cup\{a_1\})>H(D\mid Co\cup\{a_1\}\cup\{a_2\})$；…，所以 $H(D\mid C_o)>H(D\mid C_o\cup\{a_1\})>\cdots>H(D\mid C_o\cup\{a_1\}\cup\cdots\cup\{a_i\}\cup\cdots)>\cdots>H(D\mid B)$。

推论 2 说明，如果知识约简以决策表的核为起点，那么在约简中，条件熵的变化规律是单调递减。

定义 7 设 U 是一个论域，P 是 U 的一个条件属性集合，d 为决策属性，则 $Q\subseteq P$ 是 P 相对于决策属性 d 的一个约简的充分必要条件为 $H(\{d\}\mid Q)=H(\{d\}\mid P)$；且对于 Q 中任意属性 r 都有 $H(\{d\}\mid Q)\neq H(\{d\}\mid Q-\{r\})$成立。

定义 8 决策表的核属性为其所有属性约简的交集。

文献[5]根据上述定理和定义，得出了信息观下的核属性计算方法。

算法 1 决策表信息熵定义下的核属性的计算算法

输入：决策表系统 $S=\langle U,R,V,f\rangle$，$R=C\cup D$ 是属性集合，子集 $C=\{a_i\mid i=1,\cdots,m\}$和 $D=\{d\}$分别为条件属性集和决策属性集；

输出：信息熵定义下的核属性 $CORED(C)$；

第一步 $CORED(C)=\phi$；

第二步 对于条件属性集 C 中的所有属性 r，如果 $H(\{d\}\mid C)<H(\{d\}\mid C-\{r\})$，则 $CORED(C)=CORED(C)\cup\{r\}$；

第三步 结束。

4 基于属性核的规则提取算法

由文献[4]可知：在不相容决策表条件下，信息观所求得的核包含代数观所求得的核，即 $CORE_1\subseteq CORE_2$。也就是说，在不相容决策表条件下，信息观所求得的核与代数观所求得的核有两种关系：$CORE_1=CORE_2$ 和 $CORE_1\subset CORE_2$。即可把不相容决策表分为两类：一类为代数观和信息观所求的核属性相等。另一类为信息观所求的核包含代数观所求的核。

本文将通过基于核属性的规则提取算法来考察两种观点所求得的核属性对提取规则产生了什么样的影响。

算法 2

输入：决策表系统 $S=\langle U,R,V,f\rangle$，$R=C\cup D$ 是属性集合，子集 $C=\{a_i\mid i=1,\cdots,m\}$ 和 $D=\{d\}$ 分别为条件属性集和决策属性集；

输出：基于核属性的规则集；

第一步　$Rule(S)=\phi$；

第二步　用代数观和信息观分别求出决策表的核属性，并求出核属性的重要性；

第三步　在核属性集中，按照属性重要性依次提取单属性规则，然后单属性基础上按属性重要性逐渐增加属性数量，继续提取规则；如果得到的规则的支持度为 1，则删除满足规则的对象；

第四步　在全部核属性的基础上从 1 到 n 依次增加非核属性，提取规则，得到规则集 $Rule*(S)$；

第五步　对规则集进行约简，在产生同一决策的规则集中取支持度最大的一个作为规则，得到所求规则集 $Rule(S)$。

第六步　结束。

表 1　实验参数

组	条件属性个数	条件属性值域	决策属性个数	决策属性值域
1	5	(0,1)	1	(0,1)
2	5	(0,1,2,3,4)	1	(0,1)
3	12	(0,1)	1	(0,1)
4	5	(0,1)	1	(0,1,2,3,4,5)
5	12	(0,1)	1	(0,1,2,3,4,5)

为了得到具有统计意义的实验分析结果，我们进行了五组统计实验。在每组实验中，我们针对决策表的不同规模（决策表中所包含的样本记录数目）和复杂程度（决策表中的冲突样本数目）进行统计分析，每组测试 100 个决策表。表 1 给出了每组实验的参数设置情况。例如，在第 1 组实验中，决策表有 5 个条件属性，1 个决策属性，每个条件属性的取值有 2 种，即 0 和 1，决策属性的取值有 2 种，即 0 和 1。我们用算法 2 来计算这些决策表在代数观和信息观下的规则集。

5 组实验所得到的结果分别如图 1，图 2，图 3，图 4，图 5 所示。图中的横坐标为决策表序号，纵坐标为决策表规则个数。

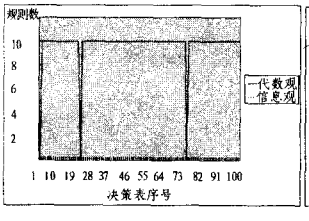

图 1　第一组试验结果

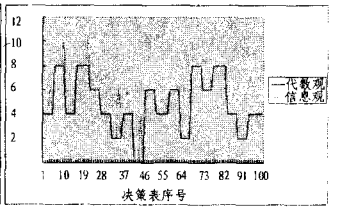

图 2　第二组试验结果

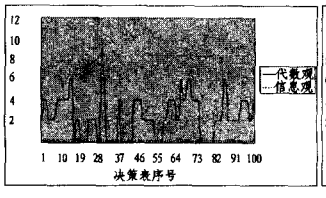

图 3　第三组试验结果

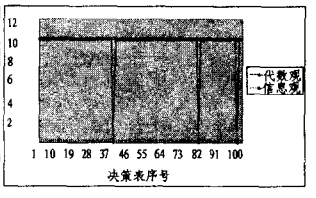

图 4　第四组试验结果

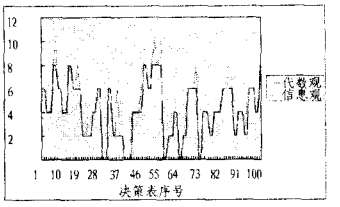

图 5　第五组试验结果

结束语　分析图 1，图 2，图 3，图 4，图 5 的结果，我们可以得到如下结论。

当不相容决策表代数观和信息观所求的核不同时（$CORE_1\subset CORE_2$），说明用信息观得到的属性重要性不为 0 的属性集比用代数观得到的属性重要性不为 0 的属性集大，用信息观得到的属性重要性为 0 的属性集将比用代数观得到的属性重要性为 0 的属性集小。设两种观点得到的核属性集相差的那部分属性组成属性集 C^-，相同的那部分属性集为 $CORE_1$。$CORE_1$ 对确定分类实例和不确定分类实例都会产生影响。C^- 在代数观下的属性重要性为 0，而在信息观下的属性重要性不为 0。说明属性集 C^- 对确定分类实例不产生影响，对不确定分类实例产生影响，减少其中的一个属性，不改变论域中本身已确定分类的实例，且所有本身不能确定分类的实例仍然不能确定分类，只是不确定性会产生变化。所以 C^- 对产生规则将产生影响，在信息观中它作为核属性优先产生规则，改变了不确定分类的不确定性，而在代数观中它作为非核属性加入到核属性中产生规则，增加了规则的复杂度，所以导致由信息观产生的规则集会比代数观得到的规则集要好（即用较少的属性产生了相同的决策）。

参 考 文 献

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

[2] 叶东毅，陈昭炯. 一个新的差别矩阵及其求核方法. 电子学报，2002，30(7)：1086-1088

[3] 王国胤，于宏，杨大春. 基于条件信息熵的决策表约简. 计算机学报，2002，25(7)：759-766

[4] Wang G Y. Algebra view and information view of rough sets theory// Proceedings of SPIE，2001，43(84)：200-207

[5] 王国胤. 决策表核属性的计算方法. 计算机学报，2003，26(5)：611-615

[6] 王国胤. Rough Set 理论代数观和信息观核属性的差异研究// 中国人工智能学会第 10 届全国学术年会论文集. 广州，2003

[7] 王国胤. Rough 集理论在不完备信息系统中的扩充. 计算机研究与发展，2002，39(10)：1238-1243

[8] Hu X H，Cercone N. Learning in relational databases：a rough set approach. Com-putational Intelligence，1995，11(2)：323-337

[9] Pawlak Z，Grzymala-Busse J，Slowinski R，et al. Rough sets. Communication of the ACM，1995，38(11)：89-95

[10] 刘宗田. 属性最小约简的增量式算法. 电子学报，1999，27(11)：96-98

[11] 石峰，娄臻亮，张永清. 一种改进的粗糙集属性约简启发式算法. 上海交通大学学报，2002，36(4)：478-481

[12] Han J W，Kamber M. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers，NIJ Journal，2001

[13] Kumar A. New Techniques for Data Reduction in a Database System for Knowledge Discovery Applications. Journal of Intelligent Information Systems，1998，10(1)：31-48