第 24 卷第 2 期
2007 年 04 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 24 No. 2
Apr. 2007

文章编号：1000−8152(2007)02−0183−04

一种基于有序属性决策系统分类规则提取策略

张文宇

(西安邮电学院 管理系, 陕西 西安 710061)

摘要: 分类规则的精度取决于分类算法的构造. 论文在综合分析基本粗糙集合概念及其约简算法的基础上, 阐述了一种基于准则的有序属性决策系统的数据挖掘算法. 为此首先介绍了基于有序属性决策系统的集合表达, 然后利用有序属性决策系统中准则集与属性集的基本特征构造上下近似扩展模型, 得到准则集决策系统的四个相关参数. 并进一步提出相应的数据约简与分类规则提取算法. 最后给出了用此算法约简有序属性决策系统的算例, 实验结果表明此方法挖掘出的规则简练, 更具合理性和可靠性.

关键词: 数据挖掘; 粗糙集; 决策系统; 准则; 分类质量

中图分类号: TP181　　**文献标识码**: A

Classification rule extracting strategy based on decision system with ordered attributes

ZHANG Wen-yu

(Department of Management, Xi'an Post and Telecommunication College, Xi'an Shaanxi 710061, China)

Abstract: The precision of classification rule is decided by the construction of classification algorithm. By the concepts and attribute reduction algorithm of basic rough set, a data mining algorithm based on the ordered character of attribute in decision system is proposed in this paper. First, the aggregation expression in decision system with ordered character of attribute is briefly introduced. Then, based on the basic characterization of criteria sets and attribute sets in decision system with ordered attributes, the upper and lower approximation expansion models are constructed to obtain the four relative parameters in decision system with ordered attributes. Thirdly, the corresponding data mining and classification rule extracting algorithm is constructed by using the proposed approach. Finally the rationality of the ordered attribute reduction method is validated by simulation example, and the result shows the rules mined by the method are concise and reliable.

Key words: data mining; rough set; decision system; criteria; classification quality

1　引言(Introduction)

数据挖掘是当前人工智能和数据库技术的研究热点之一. 目前, 有许多基于机器学习、模式识别及统计学的规则获取方法, 近年来的粗集理论也为数据挖掘提供了重要工具, 其导出的规则精炼而且便于存储和使用. 粗糙集是波兰数学家Z.Pawlak在1982[1,2]年提出的一种分析数据的数学理论, 其特点是不需要预先给定某些特征或属性的数量描述, 而是直接从给定问题的描述集合出发, 依据不同的观测点把对象集划分成等价类来确定给定问题的近似域, 从而找出该问题中的本质特征和内在规律[3].

本文在简述基本粗集基本概念的基础上, 采用属性的有序特征对决策系统进行简化进而提取规则, 并利用算例对该方法进行了合理性验证.

2　基本粗糙集(Basic rough set)

定义 1　称$S = (U, Q, \{V_q\}, f)$为知识表示系统. 其中: U为非空有限集, 称论域; Q为非空有限集, 称属性集合; V_q为属性$q \in Q$的值域; $f : U \to V_q$为一单射, 使论域U中任一元素取属性q在V_q中的某一唯一值. 如果Q由条件属性集合C和决策属性集合D组成, C, D满足$C \bigcap D = \phi$,$C \bigcup D = Q$, 则称S为决策系统.

定义 2　令 $S = (U, C \bigcup \{D\})$ 为一决策系统, $B \subseteq C$, 定义 B 的不可分辨关系 $IND(B)$ 为 $IND(B) = \{(x, x^{'}) \in U^2 | \forall a \in B, a(x) = a(x^{'})\}$, $a(x)$ 为元素x在属性a上的值. 若$(x, x^{'}) \in IND(B)$,

收稿日期: 2005−07−28; 收修改稿日期：2006−04−30.
基金项目: 陕西省教育厅专项科研计划项目基金(05JK092).

说明根据已有的信息不能将x和x'区分开.

定义 3 令集合V_d表示决策属性d的域值, 不失一般性, 设$V_d=\{1,2,\cdots,r(d)\}$, 则决策属性d将论域U划分成$\{Y_1,Y_2,\cdots,Y_{r(d)}\}$, 其中$Y_k=\{x\in U|d(x)=k,1\leqslant k\leqslant r(d)\}$, $d(x)$为决策属性d在对象x上的值, 集合Y_k称为决策系统S的第k个决策类或决策概念.

利用基本粗集理论对决策系统进行数据约简目前已经有很多的方法并在实际应用中取得了可喜的成果. 例如基于正域的约简方法、基于分辨矩阵的约简方法、基于信息熵的约简方法、基于距离函数的约简方法[4~11]等, 每种方法都具有各自的特点及应用场合. 下面在Salvatore Greco, Roman Slowinski[12]等教授提出的基于准则决策系统的数据约简的方法基础上提出带有有序属性决策系统的数据挖掘新算法, 并将其应用于电信网络质量评价中.

3 带有有序属性的粗糙近似(Rough approximation with ordered attributions)

3.1 基本原理(Basic theory)

基于有序属性的数据挖掘方法同样应用于如定义1所示的决策系统中, 但是决策系统中的属性是通过某种递增递减顺序进行排序而形成的准则域. 对于$\forall q\in C$, 且q为一准则, 则存在优先关系S_q, 使xS_qy表示“针对属性q而言, x至少与y一样好”, 则S_q是定义在U上的完备传递二元有序关系. 相反, 对于$\forall q\in C$, q不是准则而是属性, 则存在等价关系I_q, 使xI_qy为基本粗集理论中的等价二元关系, 即它满足自反性、对称性和传递性. 标记$C^{>}$为C中的准则子集, 而$C^{=}$为C中的属性子集即$C^{>}\bigcup C^{=}=C$, $C^{>}\bigcap C^{=}=\phi$. 而且对于$\forall P\subseteq C$, 标记$P^{>}$为包含在C中的准则集, 即$P^{>}=P\bigcap C^{>}$, $P^{=}$为包含在C中的属性集, 即$P^{=}=P\bigcap C^{=}$, 且满足$P=P^{>}\bigcup P^{=}$,$P^{>}\bigcap P^{=}=\phi$. 令R_p是一个U上的自反和传递二元关系, 即R_p是在U上部分有序的关系, 精确地说, 对于$\forall P\subseteq C$, 可定义R_p如下: $\forall x,y\in U, xR_py$, 当且仅当若$q\in P^{>}$则存在xS_qy关系, 若$q\in P^{=}$, 则存在xI_qy. 如果$P\subseteq C^{>}$及xR_py, 则表示对于$\forall q\in P$, x优先于y. 广义地讲, 令$CL=\{CL_t,t\in T\}$, $T=\{1,2,\cdots,n\}$, 是U上的一个划分集, 即对于$\forall x\in U$, x属于且仅属于一个$CL_t\in CL$. 假定$\forall r,s\in T$, 若$r>s$, 则CL_r的元素优于CL_s的元素. CL类的划分代表U上对象的综合评价: 最差的对象位于CL_1中, 最好的对象位于CL_n中, 其他的对象属于CL_r.

从CL划分出发, 可定义如下集合:

$$\begin{cases} CL_t^{\geqslant}=\bigcup\limits_{s\geqslant t}cl_s,\\ CL_t^{\leqslant}=\bigcup\limits_{s\leqslant t}cl_s. \end{cases}$$

对于$\forall P\subseteq C$, 令

$$\begin{cases} R_p^{+}(x)=\{y\in U: yR_px\},\\ R_p^{-}(x)=\{y\in U: xR_py\}. \end{cases}$$

对于$x\in U$, $R_p^{+}(x)$表示在准则集中优先于x的对象与在属性集P中与x等价的对象之和. 类似地$R_p^{-}(x)$表示在准则集P中被x优先的对象与在属性集P中与x等价的对象之和.

3.2 基于准则的上下近似扩展模型(Upper and lower approximation expanding model based criteria)

假设P_p为自反的, 可以说$x\in U$确切属于$CL_t^{\geqslant}$等价于$R_p^{+}(x)\subseteq CL_t^{\geqslant}$. 对于$P\subseteq C$,$t\in T$, $y\in U$可能属于$CL_t^{\geqslant}$等价于至少存在一个对象$x\in CL_t^{\geqslant}$使得在准则集P中y优先于x及在属性集P中y等价于x, 即$y\in R_p^{+}$. 因此考虑$P\subseteq C$, 属于$CL_t^{\geqslant}$的所有对象集, 确切地构成了$CL_t^{\geqslant}$的下近似, 而可能属于$CL_t^{\geqslant}$的所有对象集构成了$CL_t^{\geqslant}$的上近似.

形式描述如下: $\forall t\in T$及$\forall P\subseteq C$, 可定义针对P而言$CL_t^{\geqslant}$的下近似$\underline{P}CL_t^{\geqslant}$, 及针对$P$而言$CL_t^{\geqslant}$的上近似$\overline{P}CL_t^{\geqslant}$:

$$\begin{cases} \underline{P}CL_t^{\geqslant}=\{x\in U: R_p^{+}\subseteq CL_t^{\geqslant}\},\\ \overline{P}CL_t^{\geqslant}=\bigcup\limits_{x\in CL_t^{\geqslant}}R_p^{+}(x). \end{cases}$$

同理可得下近似$\underline{P}CL_t^{\leqslant}$与上近似$\overline{P}CL_t^{\leqslant}$的数学表示.

类似基本粗集定义, 本文提出了基于准则集决策系统的相关参数, 并将其应用在带有准则决策系统的数据约简算法中.

参数 1 对于$\forall t\in T$及$P\subseteq C$, 可定义$CL_t^{\geqslant}$与$CL_t^{\leqslant}$的分类系数为

$$\begin{cases} \alpha_p(CL_t^{\geqslant})=\mathrm{card}(\underline{P}CL_t^{\geqslant})/\mathrm{card}(U),\\ \alpha_p(CL_t^{\leqslant})=\mathrm{card}(\underline{P}CL_t^{\leqslant})/\mathrm{card}(U). \end{cases}$$

知识P关于划分CL的分类系数为: $\alpha_p(CL)=\sum\limits_{t\in T}\{\alpha_p(CL_t^{\geqslant})+\alpha_p(CL_t^{\leqslant})\}$.

参数 2 对于$t\in T$及$\forall P\subseteq C$, 可定义$CL_t^{\geqslant}$与$CL_t^{\leqslant}$的近似精度为

$$\begin{cases} \partial_p(CL_t^{\geqslant})=\mathrm{card}(\underline{P}CL_t^{\geqslant})/\mathrm{card}(\overline{P}CL_t^{\geqslant}),\\ \partial_p(CL_t^{\leqslant})=\mathrm{card}(\underline{P}CL_t^{\leqslant})/\mathrm{card}(\overline{P}CL_t^{\leqslant}). \end{cases}$$

知识P关于划分CL的近似精度为: $\partial_p(CL)=$

$\sum\limits_{t\in T}\{\partial_p(CL_t^{\geqslant})+\partial_p(CL_t^{\leqslant})\}$ 对于近似分类的不精确性, 可定义针对P而言CL 划分的近似质量简称为分类质量为

参数 3　知识P关于划分CL的分类质量为

$$\gamma_p(CL)=\frac{\sum\limits_{t\in T}\mathrm{card}(\underline{P}CL_t^{\geqslant})+\sum\limits_{t\in T}\underline{P}CL_t^{\leqslant})}{\sum\limits_{t\in T}\mathrm{card}(\overline{P}CL_t^{\geqslant})+\sum\limits_{t\in T}\overline{P}CL_t^{\leqslant})}.$$

它表示了决策表中所有P-正确分类对象与所有P-可能正确分类对象之比. 为了表示分类的相对近似, 可定义分类正域.

参数 4　知识P关于划分CL的分类正域为

$$\beta_p(CL)=\frac{\mathrm{card}(POS_p(CL))}{\mathrm{card}(U)}=\frac{\sum\limits_{t\in T}\mathrm{card}(\bigcup\limits_{X\in CL_t^{\geqslant}}\underline{P}X)+\sum\limits_{t\in T}\mathrm{card}(\bigcup\limits_{X\in CL_t^{\leqslant}}\underline{P}X)}{\mathrm{card}(U)}.$$

每一最小子集$P\subseteq C$若满足$\gamma_p(CL)=\gamma_c(CL)$ 则称为CL的一个约简, 记为RED_{CL}. 即约简前后划分CL关于整个条件属性集合的分类质量保持不变. 一个带有准则的决策表不止一个约简, 所有约简的交称为该决策表的核. 记为CORE_{CL}, $\mathrm{CORE}_{CL}=\bigcap\mathrm{RED}_{CL}$.

3.3 基于带有准则决策系统的数据约简算法(Data reduction algorithm based on decision system with criteria)

3.3.1 参数描述(Parameter description)

设决策系统中的条件属性C中有m个属性: $C_1,C_2,\cdots,C_m$, 其中m_1个为有序属性, m_2个为无序属性. 决策属性集合D的划分为$CL=\{Cl_t, t\in T\}, T=\{1,2,\cdots,n\}$. 对每一个条件属性$C_i$计算以下4个参数: $\alpha_{C_i}(CL),\partial_{C_i}(CL),\gamma_{C_i}(CL),\beta_{C_i}(CL)$, 令$H_{C_i}$和$J_{C_i}$分别为这4个参数的算数平均和几何平均, $S_{C_i}=H_{C_i}+J_{C_i}$为条件属性的重要性系数.

3.3.2 算法过程(Algorithm procedure)

Step 1　输入决策表的条件属性与决策属性, 明确条件属性中的有序属性及无序属性并区分决策类;

Step 2　分别计算各个决策类关于整个条件属性的上下近似, 边界域正域和负域;

Step 3　对于条件属性集合C, 计算$V_c(CL)$;

Step 4　令初始约简属性集C_0为空;

Step 5　分别计算各个条件属性的参数, 计算其重要性系数S_{C_i}, 并构成集合S;

Step 6　从S中取出最大值, 将对应属性加入约简属性集C_0中, 计算$V_{C_0}(CL)$, 并将S_{C_i} 从S中删除;

Step 7　如果$V_{C_0}(CL)<V_C(CL)$, 则转Step 6, 否则转Step 8;

Step 8　C_0中包含的条件属性即为条件属性集合C的一个约简.

当不考虑属性的有序特征时, 此算法便可蜕变为标准粗糙集合的属性约简算法, 从而保证了带有有序属性的粗糙集数据约简与标准粗糙集的属性约简算法的一致性.

4 算例(Example)

在表1中, 电信网络质量评价指标一般由以下6个属性描述：C1,C2,C3,C4,C5,D分别表示网络负载能力、业务范围、信息传输时延、中心局址、拥塞标志、电信网络质量.

表 1　电信网络质量决策

Table 1　Telecommunication network quality decision

论域U	C1	C2	C3	C4	C5	D
X1	高	多	长	L1	严重	好
X2	中	多	长	L1	严重	差
X3	中	多	长	L1	严重	好
X4	低	一般	长	L1	不能呼叫	差
X5	中	一般	短	L1	不能呼叫	差
X6	高	一般	短	L1	中等	好
X7	中	一般	长	L1	中等	好
X8	高	多	长	L2	中等	好
X9	中	多	长	L2	严重	好
X10	低	一般	长	L2	不能呼叫	差
X11	中	一般	短	L2	中等	好
X12	高	一般	短	L2	中等	好

由表可知, C1,C2,C3,C5是准则, 因为它的属性值是有序的, C4是一个属性, 其属性值范畴是无序的, D是决策属性, 定义了两个有序决策, 详细地讲：高>中>低, 多>一般, 短>长, 中等拥塞>严重拥塞>不能处理呼叫, 好>差, 其中>为“优于”符号.

4.1 基本粗集方法的约简结果(Reduction result based on rough set method)

由基本粗集方法中的基于正域算法, 定义cl_1 为电信网络质量差决策类, cl_2 为电信网络质量好决策类. 经过计算, 决策规则为(计算过程略)

R1: (C1=高)→(电信网络质量=好);

R2: (C1=中)∧(C4=L2)→(电信网络质量=好);

R3: (C1=中)∧(C2= 一般)→(电信网络质量=好);

R4: (C1=中)∧(C2=多)∧(C4=L1)→(电信网络质量=好)∨(电信网络质量=差);

R5: (C1=低)→(电信网络质量=差);

R6: (C5=不能处理呼叫)→(电信网络质量=差);

R7: (C1=中)∧(C3=短)∧(C4=L1)→(电信网络质量=差).

4.2 基于优先及不可分辨关系的近似约简结果(Approximation reduction result based on priority and indiscernibility relation)

定义决策类$cl_1^{\leqslant}$为电信网络质量最多为差的情况, $cl_2^{\geqslant}$表示电信网络质量至少为好的情况, 因为只有两个决策类, 因而$cl_1^{\leqslant}$=cl_1 及$cl_2^{\geqslant}$=cl_2. 计算$cl_1^{\leqslant}$及$cl_2^{\geqslant}$的C—-上下近似及边界域分别为

$$\underline{C}cl_1^{\leqslant}=\{4,10\},\ \overline{C}cl_1^{\leqslant}=\{2,3,4,5,7,10\},$$
$$Bn_C(cl_1^{\leqslant})=\{2,3,5,7\},$$
$$\underline{C}cl_2^{\geqslant}=\{1,6,8,9,11,12\},$$
$$\overline{C}cl_2^{\geqslant}=\{1,2,3,5,6,7,8,9,11,12\},$$
$$Bn_C(cl_2^{\geqslant})=\{2,3,5,7\}.$$

$cl_1^{\leqslant}$的近似精度为0.63, $cl_2^{\geqslant}$的近似精度为0.86, 分类质量为0.67. 只有一个约简即核, $Red_{cl}(c)=CORE_{cl}(c)=C_1,C_4,C_5$; 则最小决策规则为

R1: (C1=高)→(电信网络质量至少为好);

R2: (C1至少为中)∧(C4=L2)→(电信网络质量至少为好);

R3: (C1=低)→(电信网络质量至多为差);

R4: (C1=中)∧(C4=L1)→(电信网络质量=好)∨(电信网络质量=差);

R5: (C5=不能处理呼叫)→(电信网络质量=差).

基于优先及不可分辨关系的粗集方法与基于不可分辨关系的粗集方法相比:

* 规则数目减少, 规则利用较少的属性数及描述符, 更精炼;

* 属性的约简数目减少;

* 算法涵盖了属性了有序特征. 当不考虑属性的有序特征时, 此算法便可蜕变为标准粗糙集合的属性约简算法, 从而保证了带有有序属性的粗糙集数据约简与标准粗糙集的属性约简算法的一致性.

粗糙集合理论已经广泛应用于数据分析、图像处理、声音识别、决策支持分析和数据挖掘等领域, 并取得了很大成功. 本文在基本粗糙集合理论的基础上, 利用基于有序属性的数据挖掘方法, 能够快速地在决策系统中挖掘出规则, 且挖掘出的规则简练, 更具合理性和可靠性.

参考文献(References):

[1] PAWLAK Z. Rough set approach to knowledge-based decision support[J]. *European Journal of OR*, 1999, 3(27): 48 – 57.

[2] PAWLAK Z. Rough sets[J]. *Int J of Computer and Information Science*, 1982, 11(5): 341 – 356.

[3] ZIARKO W. Data-based acquisition and incremental modification of classification rules[J]. *Compuerational Intelligence*, 1995, 4(11): 357 – 370.

[4] 叶东毅, 黄翠微, 赵斌. 基于逼近精度的粗糙集属性约简算法[J]. 福州大学学报, 2000, 28(1): 8 – 10.
(YE Dongyi, HUANG Cuiwei, ZHAO Bin. An algorithm for attributions reduction in rough set based on approximation quality[J]. *Fuzhou University Journal*, 2000, 28(1): 8 – 10.)

[5] ZIARKO W. Rough sets as a methodology for data mining[C]// POLKOWSKI L, SKOWRON A. *Rough sets in Knowledge Discovery: Methods Application*. Heidelberg: Physica-Verlag, 1998: 289 – 298.

[6] 王珏. 粗糙集理论和统计学习理论[M]//陆然峥. 知识科学与计算科学. 北京: 清华大学出版社, 2003: 49 – 51.
(WANG Yu. Rough set theory and statistical learning theory[M]//LU Ranzheng. *Knowledge Science and Computing Science*. Beijing: Tsinghua University Press, 2003: 49 – 51).

[7] 张文修. 信息系统与知识发现[M]. 北京: 科学出版社, 2003.
(ZHANG Wenxing. *Information System and Knowledge Diacovery*[M]. Beijing: Science Press, 2003.

[8] MIDELFART N, KOMOROWSKI J. A rough set approach to inductive logic programming[C]//ZIARKO W, YAO Y. *Rough Sets and Current Trends in Computing - Second International Conference*. Banff, Canada: Springer Press, 2000: 190 – 198.

[9] MAHESWARI A, UMA V, SIROMONEY A, et al. The variable precision rough set inductive logic programming model and web usage graphs[J]. *New Frontiers in Artificial Intelligence–Joint JSAI 2001 Workshop Post-Proceedings*, 2001, 2253: 339 – 343.

[10] BAZAN J G, SZCZUKA M. A collection of tools for rough set computations[C]//*Proc of the 2nd Int Conf Rough Sets and Current Trends in Computing*. Banff, Canada: Springer Press, 2000: 74 – 81.

[11] ZHONG N, DONG J Z, OHSUGA S. Using background knowledge as a basis to control the rule discovery process[C]//ZIGHED D A, KOMOROWSKI J, ZYTKOW J. *Principles of Data Mining and Knowledge Discovery*. Berlin: Springer, 2001: 691 – 698.

[12] GRECO S, MATARAZZO B, SLOWINSKI R. A new rough set approach to multicriteria and multiattributeclassification[C]//*Rough Set and Current Trends in Computing*. Berlin: Springer, 1999: 57 – 61.

作者简介:

张文宇 (1973—), 女, 博士, 西安邮电学院管理系副教授, 发表科技论文近30篇, 出版专著2本, 目前研究方向为数据挖掘、智能决策与网络可靠性, E-mail: zwy888459@sina.com 或 zwy888459@gmail.com.