

不完备信息系统中决策规则的不确定性表示

李萍¹ 瞿英^{1,2} 王芳¹ 吴祈宗¹

(北京理工大学管理与经济学院 北京 100081)¹ (河北科技大学经管学院 石家庄 050018)²

摘要 将完备信息系统中决策规则的不确定性度量引入到不完备信息系统中,充分讨论了条件属性值缺失时规则的不确定性度量表现为概率区间的特性,在此基础上给出了确信度和覆盖度的近似概率值,并利用确信度计算出规则对分类的反映程度。

关键词 粗糙集,不完备信息系统,不确定性

中图法分类号 TP391 **文献标识码** A

Uncertain Knowledge Expression to the Decision Rule of Incomplete Information System

LI Ping¹ QU Ying^{1,2} WANG Fang¹ WU Qi-zong¹

(School of Management and Economic, Beijing Institute of Technology, Beijing 100081, China)¹

(School of Economic and Management, Hebei University of Science and Technology, Shijiazhuang 050018, China)²

Abstract The uncertainty measurement method of decision rules in complete information system was firstly introduced into the incomplete information system. When some of attribute values are missing, the uncertainty measurement of decision rule shows the characteristics of probability interval. On the basis of that, the certain factor coverage factor can be measured by approximate probability value. Besides this, the reflection degree of rule to the knowledge classification can be computed using the certain factor.

Keywords Rough theory, Incomplete information system, Uncertainty

1 引言

粗糙集作为一门新的数学方法被广泛应用于各个领域,在数据分析方面,已表现出强大的优势,如提供方法找出数据中隐藏的模式、评价数据的重要性、从数据中生成规则、对获取的数据结果进行直观分析等等。一个决策系统是一个信息系统,信息系统的不确定性主要是指决策表中往往包含冲突样本的情况。王^[1]指出在许多实际领域中,信息系统可利用的证据和知识是不确定的,对其不确定性可采用贝叶斯原理中的先验概率和后验概率来表示。Z. pawlak^[2,3]从粗糙集角度来描述了在一个完备信息系统中,利用决策规则的概率分布来分析一个知识系统的不确定性,在此基础上提出了支持度、强度、确信度和覆盖度4个度量值,并用确信度来表示规则对分类的反映程度。Kryszkiewicz M^[4,5]详细讨论了不完备信息系统的相关性质,并提出如何利用不完备信息系统来推导出规则,但是对如何度量其规则的不确定性并没有深入研究。而不完备的信息系统与完备信息系统相比,本身存在更多的不确定因素,其条件属性值的缺失使得对数据进行不确定性分析更加困难。本文将从不完备信息系统的完备化扩展系统入手,首先详细讨论在不完备信息系统中,规则的不确定性表达的特性,其次给出规则的不确定度量值,从而计算规

则对知识分类的反映程度。

2 不完备信息系统与决策规则

定义 1^[1] 一个知识表达系统也是一个信息系统,可用一个四元组 $S = \langle U, AT \cup \{d\}, V, f \rangle$ 来表示:

(1) $U = \{x_1, x_2, \dots, x_n\}$ 是非空有限对象的集合,也称之为论域。

(2) AT 为条件属性集, $\{d\}$ 为决策属性集。

(3) $V = \bigcup_{a \in A} V_a$ 表示属性的集合, V_a 是属性 a 的取值范围。

(4) $f = U \times A \rightarrow V$ 代表一个信息函数,它指定 U 中的每个对象 x 的属性值。

如果 $\exists a \in AT \cup \{d\}$, 使得对于某个对象而言, V_a 取值为未知或者遗失,一般用 $*$ 来表示,则该信息系统称之为不完备信息系统。如果 $d \notin AT$, $* \notin V_d$, 意味着在决策表中,只有条件属性集有值缺失,而决策属性中没有值缺失,该不完备信息系统表也称之为不完备决策表。

表1所列的不完备决策表来自于文献[4,5],该表为汽车参数性能决策表,假设其各属性的数据服从均匀分布,表中最后一列为符合该对象的记录数。为分析方便,Price, Mileage, Size, Max-Speed 分别以字母 P, M, S, X 来代替。 $V_p = \{low,$

到稿日期:2008-10-21 返修日期:2008-12-31 本文受河北省教育厅计划(2007205)资助。

李萍(1978—),女,博士,主要研究方向为决策支持系统、信息系统理论, E-mail: 20508048@bit.edu.cn; 瞿英(1973—),女,博士生,副教授,主要研究方向为数据挖掘、信息系统风险管理; 王芳(1982—),女,博士,主要研究方向为供应链契约、决策支持系统理论; 吴祈宗(1947—),男,教授,博士生导师,主要研究方向为决策支持系统、运筹学与控制论。

$high\}, V_s = \{full, compact\}, V_x = \{low, high\}, V_M = \{low, high\}, V_d = \{poor, good, excellent\}$ 。

表1 不完备决策表

Car	Price	Mileage	Size	MaX-Speed	D	NumberCounts
1	High	Low	Full	Low	Good	60
2	Low	*	Full	Low	Good	10
3	*	*	Compact	Low	Poor	5
4	High	*	Full	High	Good	6
5	*	*	Full	High	Excellent	10
6	Low	High	Full	*	Good	9

给定一个不完备信息表 $S = (U, AT, V, f)$, $S' = (U', AT', V', f')$ 称之为 S 的一个扩展当且仅当 $S' = S$, $AT' = AT$, $f'_a(x) \neq *$ 。 S' 是 S 的完备化当且仅当 S 是 S' 的扩展并且是一个完备化的信息系统。不完备信息系统 S 的规则生成借助于其扩展 S' 来生成, 如表1的一个扩展表2, 从表2中可以得到如下规则, 该决策规则也是表1的决策规则:

$(S, compact) \wedge (X, low) \rightarrow (d, poor)$ 由对象3生成,

$(S, full) \wedge (X, high) \rightarrow (d, good) \vee (d, excellent)$ 由对象6生成。

表2 扩展的不完备决策表

Car	Price	Mileage	Size	MaX-Speed	D	NumberCounts
1	High	Low	Full	Low	Good	60
2	Low	Low	Full	Low	Good	10
3	High	High	Compact	Low	Poor	5
4	High	Low	Full	High	Good	6
5	High	Low	Full	High	Excellent	10
6	Low	High	Full	High	Good	9

3 不完备信息表的不确定性知识表示

3.1 不完备信息表的不确定性性质

在完备的信息系统表中, 决策规则的不确定性分析可以用规则的支持度、强度、确信度、覆盖度来表示。不完备信息表中由于条件属性值的缺失, 对于从其扩展的信息表中推导出的决策规则, 无法确定其支持度、强度、确信度和覆盖度, 从而对整个信息系统表的不确定性更难以确定。文献[6,7]中的算法描述了如何从不完备的信息表中生成所有的规则, 表1中可以推导出以下规则:

- (1) $(S, full) \wedge (X, low) \rightarrow (d, good)$
- (2) $(S, compact) \rightarrow (d, poor)$
- (3) $(P, high) \wedge (M, high) \wedge (X, low) \rightarrow (d, poor)$
- (4) $(P, low) \wedge (M, low) \wedge (X, high) \rightarrow (d, excellent)$
- (5) $(S, full) \rightarrow (d, good)$
- (6) $(S, full) \rightarrow (d, excellent)$
- (7) $(X, high) \rightarrow (d, good)$
- (8) $(X, high) \rightarrow (d, excellent)$

对以上规则进行分析, 对于 $(S, full) \wedge (X, low) \rightarrow (d, good)$ 来说, 除了对象1,2满足该规则外, 对象6也有可能满足该规则, 则该规则的强度、支持度、确信度、覆盖度随之发生变化, 在此从区间的角度来进行讨论规则的强度和支持度、确信度、覆盖度的概率取值。

先对以下性质中的符号进行说明, $\|C\|$ 为符合某规则中条件属性集的所有对象, $card(\|C \wedge D\|^{lower})$ 为完全满足规则的对象个数, $card(\|C \wedge D\|^{upper})$ 为完全满足规则及可能满足规则的对象个数。 $card(\|C\|)$ 表示满足规则中

条件属性值的对象个数。 $card(\|D\|)$ 为满足规则中决策属性值的对象个数。

性质1 在不完备信息系统中, 对任一规则 $C \rightarrow D$, 如 $\forall x \in \|C\|, a \subseteq C$, 使得 $f_a(x) \neq *$, 其规则的支持度为固定值 $supp(C, D) = card(\|C \wedge D\|)$, 反之, 若 $\exists x \in \|C\|, a \subseteq C$, 使得 $f_a(x) = *$, 其规则的支持度为一个区间, 其区间范围为 $[card(\|C \wedge D\|^{lower}), card(\|C \wedge D\|^{upper})]$ 。

性质2 在不完备信息系统中, 对任一条规则 $C \rightarrow D$, 如 $\forall x \in \|C\|, a \subseteq D$, 使得 $f_a(x) \neq *$, 其规则强度为固定值 $sth(C, D) = \frac{supp(C, D)}{card(U)}$, 若 $\exists x \in \|C\|, a \subseteq D$, 使得 $f_a(x) = *$, 则规则强度为概率区间: $[\frac{supp(C, D)^{lower}}{card(U)}, \frac{supp(C, D)^{upper}}{card(U)}]$, $supp(C, D)^{lower}$ 和 $supp(C, D)^{upper}$ 为规则支持度的下限及上限。

性质3 在不完备信息系统中, 对任一规则 $C \rightarrow D$, 如 $\forall x \in \|C\|, a \subseteq D$, 使得 $f_a(x) \neq *$, 其规则的确信度为固定值 $cer(C, D) = \pi(D|C) = \frac{card(\|C \wedge D\|)}{card(\|C\|)}$, 反之, 若 $\exists x \in \|C\|, a \subseteq C$, 使得 $f_a(x) = *$, 其规则的确信度为一个区间 $[\frac{card(\|C \wedge D\|^{lower})}{card(\|C\|)}, \frac{card(\|C \wedge D\|^{upper})}{card(\|C\|)}]$ 。

推论1 不完备信息系统中, 任何一条确定规则的确信度 $cer(C, D) = 1$ 。

证明, 依据确定规则的定义, 规则 $C \rightarrow D$ 是确定的当且仅当 $C \rightarrow D$ 是明确的并且 $\|C\| \subseteq \|D\|$ 。 $card(\|C \wedge D\|) = card(\|C\|)$, 则 $cer(C, D) = \pi(D|C) = \frac{card(\|C \wedge D\|)}{card(\|C\|)} = 1$ 。

性质4 在不完备信息系统中, 对任一规则 $C \rightarrow D$, 如 $\forall x \in \|C\|, a \subseteq D$, 使得 $f_a(x) \neq *$, 其规则的覆盖度为固定值 $cov(C, D) = \pi(D|C) = \frac{card(\|C \wedge D\|)}{card(\|D\|)}$, 反之, 若 $\exists x \in \|C\|, a \subseteq C$, 使得 $f_a(x) = *$, 其规则的支持度为一个区间 $[\frac{card(\|C \wedge D\|^{lower})}{card(\|D\|)}, \frac{card(\|C \wedge D\|^{upper})}{card(\|D\|)}]$ 。

根据以上分析, 对表2求取各规则的支持度、强度、确信度和覆盖度的区间范围如表3所列。

表3 区间范围表

规则	支持度	强度	确信度	覆盖度
1	[70, 79]	[0.70, 0.79]	1	[0.82, 0.93]
2	5	0.02	1	1
3	[0, 5]	[0, 0.01]	[0, 1]	[0, 1]
4	[0, 10]	[0, 0.1]	[0, 1]	[0, 1]
5	85	0.85	0.89	1
6	10	0.10	0.11	1
7	[6, 15]	[0.06, 0.15]	[0.51, 1]	[0.12, 0.18]
8	10	0.10	0.49	1

3.2 规则的不确定性度量

针对以上对不完备信息系统中确信度和覆盖度的分析, 在此给出不完备信息系统中确信度和覆盖度的近似概率的定义。

定义2 在有 n 个属性的不完备信息系统中, 设满足和可能满足规则 $C \rightarrow D$ 的对象种类有 t 种, 则对每条规则 $C \rightarrow D$ 的确信度 $cer(C, D)$ 定义为:

$$cer(C, D) = \frac{\sum_{i=1}^i card(\|C \wedge D\|_i) * \prod_{j=1}^n P_{a_j}(\|C\|_i)}{\sum_{i=1}^i card(\|C\|_i) * \prod_{j=1}^n P_{a_j}(\|C\|_i)} \quad (4)$$

$$P_{a_j}(\|C\|) = \begin{cases} 1, & a_j(x) \neq *, \forall a_j \in \|C\| \\ \frac{1}{V_{a_j}}, & a_j(x) = *, \forall a_j \in \|C\| \end{cases} \quad (5)$$

同时覆盖度 $cov(C, D)$ 定义为:

$$cov(C, D) = \frac{\sum_{i=1}^i card(\|C \wedge D\|_i) * \prod_{j=1}^n P_{a_j}(\|C\|_i)}{card(\|D\|)} \quad (6)$$

a_i 为符合规则的条件属性, $P_{a_j}(\|C\|)$ 为该属性对应的概率值, $\sum_{i=1}^i card(\|C \wedge D\|_i) * \prod_{j=1}^n P_{a_j}(\|C\|_i)$ 是完全符合规则和可能符合规则的对象的个数, 对于一个对象来说, 如果它完全符合某条决策规则, 则其概率值 $P_{a_j}(\|C\|) = 1$, 如果它可能符合某规则, 则其概率值 $P_{a_j}(\|C\|)$ 为对象中缺失的条件属性值域的倒数之乘积。

例如, 对表 2 所生成的某条决策规则 $(P, high) \wedge (M, high) \wedge (X, low) \rightarrow (d, poor)$, 由于对象 3 有可能符合该条规则, 对象 3 的个数为 5, 但属性 P, M 的值缺失, 以上两个属性的值域范围均为 2, 因此, 计算其确信度和覆盖度如下:

$$\begin{aligned} Cer((P, high) \wedge (M, high) \wedge (X, low) \rightarrow (d, poor)) \\ = \frac{5 * 0.5 * 0.5}{5 * 0.5 * 0.5} = 1 \\ cov((P, high) \wedge (M, high) \wedge (X, low) \rightarrow (d, poor)) \\ = \frac{5 * 0.5 * 0.5}{5} = 0.25 \end{aligned}$$

3.3 规则对知识分类的反映程度

对于某一规则 $C \rightarrow D$ 记为 $R, X = \|R_x\|$ 为符合规则 R 中的条件属性集的对象集合, E 为条件属性全集的某个分类。如果 $X \cap E = E$, 则称规则 R 与 E 有关, 如果 $X \cap E = \emptyset$, 则称规则 R 与 E 无关^[1]。如果某规则 R 只与某一个分类有关, 则定义该规则的反映程度如下:

$$\mu_R(E) = cer(R_i) \quad (7)$$

即某规则对分类的反映程度为其确信度。如果某一分类 R 与多个规则有关, 则规则对分类的平均反映程度为:

$$\mu_{rule}(E) = \frac{\sum_{i=1}^m cer(R_i)}{m} = \frac{\sum_{i=1}^m \mu_{R_i}(E)}{m} \quad (8)$$

m 表示该条件属性全集的某个分类 E 中含有的规则数。

对于表 2 来说, 6 个对象组成 6 个分类, 对于第 6 个分类来说, 有第 5, 6, 7 和 8 条规则同它相关, 因此这些规则对该分类的平均反映程度为:

$$\begin{aligned} \mu_{rule}(E_6) &= \frac{\mu_{R_5}(E_6) + \mu_{R_6}(E_6) + \mu_{R_7}(E_6) + \mu_{R_8}(E_6)}{4} \\ &= 0.50 \end{aligned}$$

结束语 针对完备信息系统中规则的不确定性度量确定为定值的情况, 本文将不确定性的度量引入到不完备信息系统中。由于部分属性值的缺失, 对于某条规则, 无法判定那些具有缺失属性值的对象是否符合该规则。由此在不完备信息系统中, 如符合该规则的对象条件属性值无缺失, 则其支持度、强度、确信度和覆盖度为固定值, 如符合该规则的对象条件属性值有缺失, 则其不确定性度量是一个区间范围。在分析了不完备信息系统中规则的不确定表示的特性后, 提出以近似概率值来表示不完备信息系统的确信度和覆盖度, 并依据确信度的近似值来计算规则对条件属性全集分类的反映程度。对不完备信息系统整个规则集的不确定性的衡量还需要作进一步讨论。

参考文献

- [1] 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
- [2] Pawlak Z. Rough classification[J]. Human-computer Studies, 1999, 51: 369-383
- [3] Pawlak Z. Rough sets and intelligent data analysis[J]. Information Science, 2002, 147: 1-12
- [4] Pawlak Z. Rough sets, decision algorithms and Bayes' theorem[J]. European Journal of Operational Research, 2002, 136: 181-189
- [5] Greco S, Pawlak Z, Slowinski R. Can Bayesian confirmation measures be useful for rough set decision rules[J]. Engineering Application of Artificial Intelligence, 2004, 17: 345-361
- [6] Kryszkiewicz M. Rough Set Approach to Incomplete Information System[J]. Information Sciences, 1998, 112: 39-49
- [7] Kryszkiewicz M. Rules in incomplete information systems[J]. Information Science, 1998, 113: 274-292
- [8] 黄冰. 基于粗糙集的不完备信息系统知识理论获取理论与方法[M]. 南京: 南京理工大学, 2004

(上接第 126 页)

- [6] Mirkovic J, Reiher P, Fahmy S, et al. Measuring Denial of Service[C]//Proceedings of the 2nd ACM Workshop on Quality of Protection. Alexandria VA, USA, 2006
- [7] Pamula J, Jajodia S, Ammann P, et al. A Weakest-adversary Security Metric for Network Configuration Security Analysis[C]//Proceedings of the 2nd ACM Workshop on Quality of Protection. Alexandria VA, USA, 2006
- [8] Wang Lingyu, Singhal P A, Jajodia P S. Toward measuring network security using attack graphs[C]//Proceedings of the 2nd ACM Workshop on Quality of Protection. Alexandria VA, USA, 2007
- [9] Ozment A. Improving vulnerability discovery models[C]//Proceedings of the 2nd ACM Workshop on Quality of Protection. Alexandria VA, USA, 2007
- [10] Abedin M, Nessa S, Al-Shaer E, et al. Vulnerability analysis for

evaluating quality of protection of security policies[C]//Proceedings of the 2nd ACM Workshop on Quality of Protection. Alexandria VA, USA, 2006

- [11] Payne S C. A Guide to Security Metrics [EB/OL]. http://www.sans.org/reading_room/whitepapers/auditing/55.php, 2008-07-25
- [12] 闫强, 陈钟, 段云所, 等. 信息系统安全度量与评估模型[J]. 电子学报, 2003, 31(9): 1351-1355
- [13] 刘学忠, 刘增良, 余达太. 基于 AHP 度量模型的安全管理度量方法[J]. 微计算机信息, 2007, 23(18): 33-34
- [14] SSE-CMM Author Group. SSE-CMM (V2.06) [M]. SSE-CMM Author Group, 1999
- [15] McHugh J. Quality of Protection: Measuring the Unmeasurable[C]//Proceedings of the 2nd ACM Workshop on Quality of Protection. Alexandria VA, USA, 2006
- [16] 王涛, 郭荷清, 姚松涛. 基于综合安全保护质量的安全服务协商模型[J]. 计算机工程与科学, 2006, 28(4): 26-29