文章编号:2095-2163(2019)05-0067-04 中图分类号:TP18 文献标志码:A

完备决策信息表中一种新的相容性度量

杜 艳
(西安石油大学 计算机学院,西安 710065)

摘 要: 如何衡量决策信息表中的相容性是粗糙集领域的一个重要研究课题,这对于科学、合理的决策具有重要意义。本文指出了已有几种决策信息表相容度量方法的不足之处,在此基础之上,提出了一种新的相容性度量方法。详细研究了该度量方法的性质,并通过实例说明该方法较其它方法的合理之处。

关键词: 决策信息表;相容性;决策规则

On consistency measure of complete decision tables

DU Yan
(School of Computer Science, Xi'an Shiyou university, Xi'an 710065, China)

【Abstract】 How to measure the consistency of decision tables has become an important issue in the study of Rough Set theory, which is of great significance in the scientific decision making. By pointing out some limitations the existing consistency measures have, the paper proposes a new approach to measuring the consistency in decision tables. The properties of the proposed approach is investigated in detail, also, an example is employed to illustrate the superiority over other approaches.

【Key words】 decision tables; consistency; decision rule

0 引 言

粗糙集[1]是由波兰数学家 Pawlak 提出的一种用于处理不确定性、不完备信息的数学工具,其基本思想是在保持分类能力不变的情况下,通过知识约简,导出问题的决策和分类规则。经过近 30 年的发展,粗糙集无论是在理论研究、还是实际应用方面都取得了长足的进步,展现出可观的应用前景。目前,粗糙集理论已广泛应用于模式识别、图像处理、特征提取、神经计算、冲突分析,数据挖掘以及知识发现等领域[2-5]。

决策表[3]是一个特殊的具有条件属性和决策属性的知识表达系统,在决策信息表中,可以导出具有如下形式的决策规则:$r_{ij}: X_i \to Y_j, X_i \in U/C, Y_j \in U/D, X_i \cap Y_j \neq \varnothing$。根据决策表所提供的信息,决策表可细分为相容决策信息表和不相容决策信息表两大类。具体来说,在相容决策信息表中,每条决策规则都是完全确定性的,或者说其确定性因子为 1。在不相容决策表中,情形与此大不相同,对于任一决策规则 $r_{ij}: X_i \to Y_j$,其确定性因子不再是简单、确定的 1,而是取自(0,1)中的实数。故而,对于一个决策信息表,其相容与否对基于该信息表的决策至关重要,自然地,通过对决策规则的不确定性度量,可以导出对决策信息表的相容性度量。在本文中,正是以此为出发点,对决策信息表中的相容性进行研究。与已有文献不同的是,本文通过对不同条件属性等价类的相容度进行带有权重的聚合,提出了一种新的条件属性等价类关于决策信息的决策相容度,并结合实例分析了该方法较已有方法的有效性。对此拟展开研究论述如下。

1 决策信息表及其相容性度量

1.1 决策信息表

定义 1 一个知识表达系统是一个四元组 $S = (U, A, V, f)$,其中,$U = \{u_1, u_2, \cdots, u_{|U|}\}$ 是一个由研究对象组成的非空有限集合,称之为论域;A 是由若干属性组成的非空有限集合;$V = \cup_{a \in A} V_a$,这里 V_a 是属性 a 的值域;$f: U \times A \to V$ 是一个映射,也称为信息函数,可为每个对象的每个属性赋予一个信息值,即 $\forall x \in U, a \in A, f(x,a) \in V_a$。

在知识表达系统中,任给属性集 $P \subseteq A$,可定义基于 P 的不可区分关系 $ind(P)$ 如下:

$$ind(P) = \{(x,y) \in U \times U \mid \forall a \in P, f(a,x) = f(a,y)\}, \quad (1)$$

如果 $(x,y) \in ind(P)$,则称 x 和 y 是 P 不可区分的。容易证明不可区分关系 $ind(P)$ 是 U 上的一

作者简介:杜 艳(1994-),女,硕士研究生,主要研究方向:大数据、粗糙集。

收稿日期:2019-07-13

等价关系,以下用符号 $U/ind(P)$ (或 U/P)表示不可区分关系 $ind(P)$ 在 U 上导出的划分(或者称由属性集 P 所导出的划分), $ind(P)$ 中的等价类称为 P 基本集,符号 $[x]_P$ 表示包含 $x \in U$ 的 P 基本集。

定义 2 决策信息表是具有条件属性和决策属性的知识表达系统 $S=(U,A,V,f)$,其中 $A=C\cup D$, $C\cap D=\varnothing$, C 称为条件属性集, D 称为决策属性集。

在决策信息表中,用 $des(X_i)$ 表示对条件属性等价类的描述,即等价类 X_i 对于各条件属性值的特定取值, $des(Y_j)$ 表示对等价类 Y_j 的描述,即等价类 Y_j 对于各决策属性值的特定取值。

在决策信息表中,可导出如下形式的决策规则:

$$Z_{ij}: des(X_i)\rightarrow des(Y_j), X_i\cap Y_j\neq\varnothing. \quad (2)$$

规则 Z_{ij} 的确定性因子定义为 $\mu(Z_{ij})=\frac{|X_i\cap Y_j|}{|X_i|}$,显然, $0\leqslant\mu(Z_{ij})\leqslant 1$,当 $\mu(Z_{ij})=1$ 时,规则 Z_{ij} 是确定的,当 $0<\mu(Z_{ij})<1$ 时,称 Z_{ij} 是不确定的。以下记:

$$Rule_i=\{X_i\rightarrow Y_j \mid Y_j\in U/D, X_i\cap Y_j\neq\varnothing\}$$

1.2 已有的决策信息表中的相容性度量

已有的决策信息表中的相容性度量方法主要包括 Pawlak 于文献[2]中提出的相容性度量方法(以下简称为 Pawlak 方法)以及钱宇华于文献[6-7]提出的相容性度量方法(以下简称为钱方法),以下将要看到,文献[6-7]中提出的度量方法实际上是文献[2]中所提方法的改进形式。

定义 3 Pawlak **方法** 设 $S=(U,C\cup D,V,f)$ 是一决策信息表, $F=(Y_1,Y_2,\cdots,Y_n)$ 是由决策属性集 D 所导出的论域 U 上的划分,定义决策信息表 S 的相容度为:

$$c_C(D)=\frac{\sum_{i=1}^{n}|\underline{C}Y_i|}{|U|}, \quad (3)$$

其中, $\underline{C}Y_i=\cup\{x\in U\mid [x]_C\subseteq Y_i\in F\}$ 。

该方法的主要思想是通过衡量用条件属性集 C 能正确划分到决策等价类中的元素集在整个论域中所占的比重进而确定决策信息表的相容程度。然而,正如文献[6-7]中所述,当相容度为 0 时,Pawlak 方法并不能很好地刻画决策信息表的相容性。由定义 3 不难看出,若相容度为 0,任意决策等价类关于条件属性集 C 的粗糙下近似集为空集,或者说,论域 U 中没有一个元素能以确定性因子 1 划分到相应的决策等价类中,然而,或许有元素可能以确定性因子为 0.8 的程度包含于相应决策等价类,此时,按照定义 3,若完全忽略这部分元素的信息,自然不能详细刻画决策信息表相容的程度。鉴于此,文献[6-7]在上述分析的基础之上,提出了一种新的决策信息表的相容性度量方法。

定义 4 **钱方法** 设 $S=(U,C\cup D,V,f)$ 是一决策信息表, $F=(X_1,X_2,\cdots,X_m)$, $G=(Y_1,Y_2,\cdots,Y_n)$ 分别是由条件属性集 C 与决策属性集 D 所导出的论域 U 上的划分,定义决策信息表 S 的相容度为:

$$\alpha(S)=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\Big[1-\frac{4}{|X_i|}\sum_{j=1}^{N_i}|X_i\cap Y_j|\mu(Z_{ij})(1-\mu(Z_{ij}))\Big], \quad (4)$$

这里, $N_i=|Y_j\mid X_i\cap Y_j\neq\varnothing|$ 。

命题 1[6] (1)若每条规则 Z_{ij} 的确定性因子为 1,即 $\mu(Z_{ij})=1$,则 $\alpha(S)$ 能达到最大值 1。

(2)若每条规则 Z_{ij} 的确定性因子为 $\frac{1}{2}$,即 $\mu(Z_{ij})=\frac{1}{2}$,则 $\alpha(S)$ 能达到最小值 0。

(3)任一相容决策信息表 S 的 α 相容度为 1,即 $\alpha(S)=1$ 。

钱宇华在文献[7]中给出另外一种形式的相容度,其定义如下。

定义 5 **钱方法** 设 $S=(U,C\cup D,V,f)$ 是一决策信息表, $F=(X_1,X_2,\cdots,X_m)$, $G=(Y_1,Y_2,\cdots,Y_n)$ 分别是由条件属性集 C 与决策属性集 D 所导出的论域 U 上的划分,定义决策信息表 S 的相容度为:

$$\beta(S)=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\Big(1-\frac{4}{|U|}\sum_{j=1}^{|U|}\delta_{X_i}(u_j)(1-\delta_{X_i}(u_j))\Big) \quad (5)$$

这里, $\delta_{X_i}(u_j)=\frac{|X_i\cap|u_j||}{|X_i|}$ 。

命题 2[7] (1)任一相容决策信息表 S 的 β 相容度为 1,即 $\beta(S)=1$,

(2)若 $\beta(S)=1$,则 S 是反向不相容的。

由定义 4 与定义 5 不难看出,即使一个决策信息表是完全不相容的(即每个条件等价类关于决策等价类都不是相容的),但通过求得 $\mu(Z_{ij})$ 与 $\delta_{X_i}(u_j)$,并进行适当形式的聚合,也可求得相应的相容度,此时,决策信息表的相容度未必是 0。因而这样就能详细地刻画了决策信息表的相容性,从而

在一定程度上克服了 Pawlak 提出的相容度的缺陷。

值得注意的是，对于定义 4，由命题 1(2)可知，若每条规则 Z_{ij} 的确定性因子为$\frac{1}{2}$，即 $\mu(Z_{ij})=\frac{1}{2}$，则 $\alpha(S)$ 能达到最小值 0。同样，对于定义 5，不难验证，对于每个 $\delta_{X_i}(u_j)$，当 $\delta_{X_i}(u_j)=\frac{1}{2}$时，$\beta(S)$ 能取到最小值，或者说 S 是最不相容的。由 Z_{ij} 与 $\delta_{X_i}(u_j)$ 的定义容易知道，在一个决策信息表中，若与条件属性等价类 X_i 相交非空的决策属性等价类 Y_j 的个数大于 2，则决策信息表的相容度是严格大于 0 的，也就是说，满足此简单条件的决策信息表是不会最不相容的。

再者，根据前面所述，在一个决策信息表中，对于条件属性等价类 X_i，若 $|X_i|=n$，且 $\forall Y_j\in U/D$，$X_i\cap Y_j\neq\varnothing$，$\frac{|X_i\cap Y_j|}{X_i}=\frac{1}{n}$，此时依据 X_i 做出决策是最困难的，因为对任一 Y_j，X_i 中有相同个数的对象支持 Y_j，或者说所有的决策规则 $Z_{ij}:des(X_i)\rightarrow$ $\mathrm{des}(Y_j)$，$X_i\cap Y_j\neq\varnothing$，都是等概率的，从依据做出决策规则的难易程度来判断决策信息表相容性的角度来看，在此种情形下，决策表是最不相容的，或者说决策表的相容度为 0。而在前面所述的方法中，只有当与 X_i 相交非空的决策等价类的个数为 2 个，且 X_i 与任一 Y_j 相交部分在 X_i 中所占比重为$\frac{1}{2}$时，决策表的相容度才为 0，这显然有一定的局限性。本文正是以此为出发点，提出了一种新的决策信息表的相容度。

1.3　一种新的决策信息表中的相容性度量

定义 6　设 $S=(U,C\cup D,V,f)$ 是一决策信息表，$G=(Y_1,Y_2,\cdots,Y_n)$ 是由决策属性集 D 所导出的论域 U 上的划分，$\forall X_i\in U/C$，定义 X_i 关于决策属性集 D 的相容度为：

$$\gamma_i^D=\begin{cases}1, & n=1;\\ \dfrac{\max\mu(Z_{ij})-1/n}{1-1/n}, & n>1.\end{cases}\tag{6}$$

这里，$Z_{ij}=X_i\rightarrow Y_j,X_i\in U/C,Y_j\in U/D$，$X_i\cap Y_j\neq\varnothing,\mu(Z_{ij})=\dfrac{|X_i\cap Y_j|}{|X_i|}$，$n=\big|\{Y_j\in U/D\,|\,X_i\cap Y_j\neq\varnothing\}\big|$。

命题 3　(1) $0\leqslant\gamma_i^D\leqslant 1$，

(2) $\gamma_i^D=1$ 当且仅当存在 $Y_j\in U/D$，使得 $X_i\subseteq Y_j$，

(3) $\gamma_i^D<1$ 当且仅当 $n\geqslant 2$，

(4) $\gamma_i^D=0$ 当且仅当 $n\geqslant 2$ 且 $\mu(Z_{ij})=\frac{1}{k_i}$，这里 $k_i=\big|\{Y_j\in U/D\,|\,X_i\cap Y_j\neq\varnothing\}\big|$。

证明：(1)显然成立。

(2)若 $\gamma_i^D=1$，由定义 6 知 $n=1$，反之，若 $n>1$，则不难验证 $\mu(Z_{ij})<1$，从而 $\gamma_i^D<1$，矛盾！由 $n=1$ 容易推得结论自然成立。

反过来，若存在 $Y_j\in U/D$，使得 $X_i\subseteq Y_j$，则 $n=1$，由定义 6 可知 $\gamma_i^D=1$。

(3)由(1)可直接推出。

(4)若 $\gamma_i^D=0$，由(2)知 $n\geqslant 2$，再由定义 6 可知 $\max\mu(Z_{ij})-1/n=0$，从而 $\forall Z_{ij}\in Rule_i,\mu(Z_{ij})=\frac{1}{n}$。反过来，若 $n\geqslant 2$ 且 $\mu(Z_{ij})=\frac{1}{n}$，则 $\gamma_i^D=0$ 可由定义 6 立即推得。

定义 7　设 $S=(U,C\cup D,V,f)$ 是一决策信息表，$F=(X_1,X_2,\cdots,X_m)$，$G=(Y_1,Y_2,\cdots,Y_n)$ 是分别由条件属性集 C 与决策属性集 D 所导出的论域 U 上的划分，定义决策信息表 S 的相容度为：

$$C(S)=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\gamma_i^D.\tag{7}$$

这里，γ_i^D 是 X_i 关于决策属性集 D 的相容度，见定义 6。

命题 4　(1) $0\leqslant C(S)\leqslant 1$，

(2)决策信息表 S 相容当且仅当 $C(S)=1$，

(3) $C(S)=0$ 当且仅当 $\forall Z_{ij}\in Rule_i,\mu(Z_{ij})=\frac{1}{k_i}$，这里，$k_i=\big|\{Y_j\in U/D\,|\,X_i\cap Y_j\neq\varnothing\}\big|$。

证明：(1)由命题 3(1)及定义 7 可立即推得。

(2)必要性：若决策表 S 是相容的，由相容的定义可知 $\forall X_i\in U/C$，存在 $Y_j\in U/D$，使得 $X_i\subseteq Y_j$，从而由定义 6 及命题 3(2)知 $\gamma_i^D=1$，再由定义 7 可得 $C(S)=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\gamma_i^D=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\times 1=\sum_{i=1}^{m}\frac{|X_i|}{|U|}=1$。

充分性：如果 $C(S)=1$，则由定义 7 及命题 4(1)易知 $\forall X_i\in U/C$，X_i 关于决策属性集 D 的相容度 $\gamma_i^D=1$，再由命题 3(1)知，$Y_j\in U/D$，使得 $X_i\subseteq Y_j$，由 X_i 的任意性便知决策信息表 S 是相容的。

(3)由定义 7 知 $C(S)=0$ 当且仅当 $\forall X_i\in U/D$，$\gamma_i^D=0$，再由命题 3(4)知，当且仅当 $\mu(Z_{ij})=\frac{1}{k_i}$，这里 $k_i=\big|\{Y_j\in U/D\,|\,X_i\cap Y_j\neq\varnothing\}\big|$。

例1 一关于病人的决策信息表见表1。

表1 关于病人的决策信息表

Tab.1 Decision information table for patients

病人	条件属性		决策属性	
	头痛	肌肉痛	体温	流感
e_1	是	是	正常	否
e_2	是	是	高	是
e_3	是	是	很高	是
e_4	否	是	很高	否
e_5	否	否	高	否
e_6	否	是	很高	是
e_7	否	否	高	是
e_8	否	是	很高	否

显然,该决策信息表格是一完备的, $C=\{$头痛,肌肉痛,体温$\}$, $D=\{$流感$\}$,则 $U/C=\{\{e_1\},\{e_2\},\{e_3\},\{e_5,e_7\},\{e_4,e_6,e_8\}\}$, $U/D=\{\{e_1,e_4,e_5,e_8\},\{e_2,e_3,e_6,e_7\}\}$,

由于 $[e_4]_C \not\subset [e_4]_D$, $[e_5]_C \not\subset [e_5]_D$,该决策信息表是不相容的。

可求得: $\gamma_1^D=\gamma_2^D=\gamma_3^D=1$, $\gamma_5^D=\gamma_7^D=0$, $\gamma_4^D=\gamma_6^D=\gamma_8^D=\frac{1}{3}$,故由定义7知:

$$C(S)=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\gamma_i^D=\frac{1}{8}\times 1+\frac{1}{8}\times 1+\frac{1}{8}\times 1+\frac{1}{8}\times\frac{1}{3}+\frac{1}{8}\times\frac{1}{3}+\frac{1}{8}\times\frac{1}{3}+\frac{1}{8}\times 0+\frac{1}{8}\times 0=\frac{1}{2}。$$

2 结束语

在本文中,研究提出了一种新的条件属性等价类关于决策信息的决策相容度,并通过实例分析了该方法较已有方法的有效之处。仍需指出的是,本文只是针对完备信息表实现的,未来工作中,可向不完备信息表开展类似的研究,同时也可向其它不同类型的信息表(如集值信息表,模糊信息表等)拓展。此外,与不同类型信息熵的比较研究以及基于不同数据集的实验验证也有待在下一步加大研究投入力度。

参考文献

[1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Science ,1982,11(8):341-356.

[2] PAWLAK Z. Rough sets: Theoretical aspects of reasoning about data[M]. Dordrecht:Kluwer Academic Publisher, 1991.

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

[4] 梁吉业,李德玉. 信息系统中的不确定性与知识获取[M]. 北京:科学出版社, 2005.

[5] 苗夺谦, 李道国. 粗糙集理论算法与应用[M]. 北京:清华大学出版社,2008.

[6] QIAN Yuhua, LIANG Jiye, LI Deyu, et al. Measures for evaluating the decision performance of a decision table in rough set theory[J]. Information Sciences, 2008,178(1-2): 181-202.

[7] QIAN Yuhua, LIANG Jiye, DANG Chuangyin. Consistency measure, inclusion degree and fuzzy measure in decision tables [J]. Fuzzy Sets and Systems, 2008,159(18): 2353-2377.

(上接第66页)

[6] FALK K, SEDLMEIER F, JOLY L, et al. Molecular origin of fast water transport in carbon nanotube membranes: Superlubricity versus curvature dependent friction[J]. Nano letters, 2010, 10(10): 4067-4073.

[7] LIU Yingchun, WANG Qi. Transport behavior of water confined in carbon nanotubes[J]. Physical review B, 2005, 72(8): 085420.

[8] JOSEPH S, ALURU N R. Why are carbon nanotubes fast transporters of water [J]. Nano letters, 2008, 8(2): 452-458.

[9] 张凯旺,钟建新. 缺陷对单壁碳纳米管熔化与预熔化的影响[J]. 物理学报, 2008,57(6): 3679-3683.

[10] 辛浩, 韩强, 姚小虎. 单、双原子空位缺陷对扶手椅型单层碳纳米管屈曲性能的不同影响[J]. 物理学报,2008, 57(7):4391-4396.

[11] BERENDSEN H J C, GRIGERA J R, STRAATSMA T P. The missing term in effective pair potentials[J]. Journal of Physical Chemistry, 1987, 91(24): 6269-6271.

[12] DUAN Yong, WU Chun, CHOWDHURY S, et al. A point - charge force field for molecular mechanics simulations of proteins based on condensed - phase quantum mechanical calculations [J]. Journal of computational chemistry, 2003, 24(16): 1999-2012.

[13] BERENDSEN H J C, POSTMA J P M, VAN GUNSTEREN W F, et al. Molecular dynamics with coupling to an external bath [J]. The Journal of chemical physics, 1984, 81(8): 3684-3690.

[14] GOLDSMITH J, MARTENS C C. Pressure-induced water flow through model nanopores [J]. Physical Chemistry Chemical Physics, 2009, 11(3): 528-533.

[15] 刘华,李春艳,陈建超,等. 受限于碳纳米管中水分子微观结构的分子动力学模拟研究[J]. 当代化工, 2011, 40(6):625-627,658.

[16] XIONG Wei, LIU Zhe, MA Ming, et al. Strain engineering water transport in graphene nanochannels[J]. Physical Review E, 2011, 84(5): 056329.

[17] MOSADDEGHI H, ALAVI S, KOWSARI M H, et al. Simulations of structural and dynamic anisotropy in nano-confined water between parallel graphite plates[J]. The Journal of chemical physics, 2012, 137(18): 184703.