粗糙集条件信息熵权重确定方法的改进

朱红灿，陈能华
（湘潭大学 公共管理学院，湖南 湘潭 411105）

摘 要：权重确定是决策和评价的重要环节。文献[1]基于粗糙集条件信息熵的权重确定方法是为了避免冗余属性权重为0的情况，但是该方法得到的冗余属性的重要度却高于非冗余属性的重要度。通过对文献[1]粗糙集条件信息熵权重确定方法的分析，找出相矛盾的原因，并讨论了属性重要度优先级队列的构造，进而在此基础上提出了以粗糙集的代数理论为基础的粗糙集条件信息熵权重确定的改进方法。

关键词：权重；粗糙集；属性重要度；信息熵
中图分类号：O236 文献标识码：A 文章编号：1002-6487(2011)08-0154-03

0 引言

在决策过程中，通常需要建立一套完整的评价指标体系，以期对决策对象进行合理的评价，因此，需要确定各个评价指标相应的权重。指标权重反映了各个因素在评判和决策过程中所占有的地位和所起的作用，指标权重的确定关系到方案排序结果的可靠性和正确性[1]。

粗糙集理论是一种无需先验信息，处理不确定、不精确数据的数学工具[2]。粗糙集广泛应用于决策支持、专家系统、模式识别等领域[3-6]。基于粗糙集的权重确定方法主要是为了克服权重确定过分依赖专家经验知识的不足[1]，许多学者从不同角度提出了粗糙集属性重要度的确定方法，这些方法大体上可以分为三大类：第一类是基于代数理论的[3-7,9]；第二类是基于信息熵理论的[1,8]；第三类是基于区分矩阵频率函数计算的[10]。

由于数据是客观的，有时可能某些指标（属性）对于当前的数据集而言，是不太重要的，[即对应权重较小的指标(属性)]，为了更清晰地分析数据，可以将不重要的指标去掉，这是可行的，但若认为该指标的权重为0，则不太符合直觉[1]。基于此，文献[1]提出一种比较典型的基于信息熵理论的权重确定方法，能有效地避免冗余属性权重为0的情况，但是，这种权重确定方法使某些原为冗余属性的指标权重比原为非冗余属性的权重还要高，这与基于代数理论的属性重要度相矛盾，与文献[1]本身的研究基础之一——冗余属性[即对应权重较小的指标(属性)]也相矛盾。这种属性重要度错位主要是信息论观点本身的缺陷所造成的，因此，本文在寻找这些缺陷的基础上，将结合粗糙集的代数观点和信息熵理论观点，对文献[1]的粗糙集条件信息熵权重确定方法加以改进。改进可以在避免原为冗余属性的指标权重为0的基础上，解决原为冗余属性的指标的重要程度与原为非冗余属性的指标重要程度相矛盾的问题，比原有方法更加全面、合理，最后用算例说明了改进方法的有效性。

1 基本概念及原有粗糙集属性权重确定方法分析

在粗糙集理论中，条件属性(指标）的重要度是指增加某一属性给分类结果带来的变化情况，如果变化小，则该属性重要性程度低，否则，该属性重要性程度高。

基于代数表示的粗糙集理论中的概念和运算都是通过代数学的等价关系和集合运算来定义的，一般来说，其概念与运算的直观性较差，人们不容易理解其本质。基于代数表示的粗糙集权重确定相关的定义如下：

定义1[2] 在决策表$S=(U,C,D,V,f)$中，决策属性D对条件属性集$B\subseteq C$的依赖度$\gamma_B(D)=|POS_B(D)|/|U|$。

定义2[11] 在决策表$S=(U,C,D,V,f)$中，$c\in C$，则条件属性（指标）c的重要度定义为：$Sig(c)=\gamma C(D)-\gamma_{C-[c]}(D)$；条件属性$\forall c$的权重W(c)定义为：$W(c)=\frac{Sig(c)}{\sum_{a\in C}Sig(a)}$。

由定义2可知，Sig(c)越大，表示条件属性(指标)c重要程度越高，其权重也越大。Sig(c)为0，条件属性(指标)c是冗余的。

基于信息表示的粗糙集理论中的知识是从信息熵的角度来定义的，信息熵从信息的不确定性和概率测度的角度来表征信源的不定度。文献[1]中基于条件信息熵的粗糙集属性权重确定的相关定义如下：

定义3[1] 在决策表$S=(U,C,D,V,f)$中，可以认为U上任一属性集合$S\subseteq C\cup D$是定义在U上的子集组成的代数上的一个随机变量，其概率分布可以通过如下方法来确定：

基金项目：教育部人文社科研究青年基金资助项目(10YJC630421)

$[S:p]=\begin{bmatrix} S_1 & S_2 & \cdots & S_t \\ p(S_1) & p(S_2) & \cdots & p(S_t) \end{bmatrix}$，其中 $p(S_j)=|S_j|/|U|,j=1,2,\cdots,t$。

定义 4[1] 在决策表 $S=(U,C,D,V,f)$中，决策属性集 $D(U/D=\{D_1,D_2,\cdots,D_k\})$对条件属性集 $C(U/C=\{C_1,C_2,\cdots,C_m\})$的条件信息熵 $I(D|C)$定义为：

$$I(D|C)=\sum_{i=1}^{m}\frac{|C_i|^2}{|U|^2}\sum_{j=1}^{k}\frac{|D_j\cap C_i|}{|C_i|}\left(1-\frac{|D_j\cap C_i|}{|C_i|}\right)。$$

定义 5[1] 在决策表 $S=(U,C,D,V,f)$中，$\forall c\in C$，则条件属性（指标）c 的重要度定义为：$NewSig(c)=I(D|(C-\{c\}))-I(D|C)+I(D|\{c\})$；条件属性 $\forall c$ 的权重 $W(c)$定义为：

$$W(c)=\frac{NewSig(a)}{\sum_{a\in C}NewSig(a)}。$$

由定义 5 可知，$NewSig(c)$考察了属性在属性集以及属性自身的重要程度，同时可以使属性集的各属性的权重不为 0[1]。

2 条件信息熵权重确定方法改进

以上的条件信息熵权重确定方法，可以避免基于代数论的粗糙集冗余属性权重为 0 的情况，但某些原为冗余属性的指标重要度有可能比原为非冗余属性的指标更高。为了更好地说明该定义的不足，举例如表 1。

由定义 1 和定义 2 可以得出：

$Sig(c1)=\gamma_C(D)-\gamma_{C-\{c1\}}(D)=\frac{5}{10}-\frac{4}{10}=\frac{1}{10}$；$Sig(c2)=\frac{2}{10}$；$Sig(c3)=Sig(c4)=0$；

$W(c1)=\frac{1/10}{1/10+2/10}=0.33$；$W(c2)=0.67$；$W(c3)=W(c4)=0$。

由以上计算结果可以看出，属性 c3,c4 是冗余属性，属性权重为 0，属性 c2,c1 的重要性程度更高，在决策过程中所起决定性作用，得到的指标权重顺序是：$W(c2)>W(c1)>W(c3),W(c4)$。

由定义 4 和定义 5 可以得出：

$NewSig(c1)=I(D|(C-\{c1\}))-I(D|C)+I(D|\{c1\})=\frac{8}{100}-\frac{6}{100}+\frac{14}{100}=\frac{16}{100}$；

$NewSig(c2)=\frac{28}{100}$；$NewSig(c3)=\frac{26}{100}$；$NewSig(c4)=\frac{14}{100}$；

$W(c1)=\frac{16/100}{16/100+28/100+26/100+14/100}=0.19$；$W(c2)=0.33$；$W(c3)=0.31$；$W(c4)=0.17$。

由以上计算结果可以看出，原来为冗余属性 c3,c4 的权重都不为 0，得到的指标权重顺序是：$W(c2)>W(c3)>W(c1)>W(c4)$，可以看出，c3 的权重却高于原为非冗余属性的 c1，其重要性程度更高，这种结果与基于代数表示的粗糙集属性重要度相矛盾。

基于以上分析，本文给出一个以粗糙集的代数理论为基础的粗糙集条件信息熵权重确定的改进方法，一方面，按照基于代数表示的属性重要度计算，将重要度不为 0 的非冗余属性和重要度为 0 的冗余属性分列高级优先级队列和低级优先级队列中，高级优先级队列属性集中所有的属性重要度都高于低级优先级队列属性集中的任一属性的重要度。表 1 决策表中高、低级优先队列的属性情况如下：

高级优先队列：c1，c2；

低级优先队列：c3，c4。

由于高级优先队列的任一属性的重要度高于低级优先队列的任一属性的重要度，所以，无论原为冗余属性的指标的条件信息熵的重要度是否大于原为非冗余属性的指标的条件信息熵的重要度，其重要度都低于原非冗余属性的指标，这样基于代数表示的权重方法确定的重要度较高的属性依然较为重要。另一方面，按照文献[1]的粗糙集条件信息熵的属性重要度计算，可以避免冗余属性权重为 0 这种不实际的情况。在定义 5 的属性（指标）c 的重要度 $NewSig(c)$的基础上，下面给出改进权重确立方法的定义。

定义 6 在决策表 $S=(U,C,D,V,f)$中，$\forall c\in C$，则条件属性（指标）c 的优先度(c)定义为：

$$\mu(c)=\begin{cases}\max\limits_{a\in\{x|x\in C.Sig(x)=0\}}(NewSig(a)) & Sig(c)>0\\ 0 & Sig(c)=0\end{cases}。$$

由定义 6 可知，$\mu(c)$为 0，表示条件属性 c 是冗余属性，列入低级优先级队列，列入高级优先级队列的非冗余属性 c 的优先度 $\mu(c)$为低级优先级队列的所有属性中的条件属性（指标）的基于条件信息熵的重要度的最大值。

定义 7 在决策表 $S=(U,C,D,V,f)$中，$\forall c\in C$，则条件属性（指标）c 的权重 $NewW(c)$定义为：

$$NewW(c)=\frac{NewSig(C)+\mu(c)}{\sum_{a\in C}NewSig(C)+\mu(a)}。$$

在该定义中，$NewSig(c)$表示基于信息表示的条件属性 c 的重要程度，$\mu(c)$则表示基于代数表示的条件属性 c 所处的队列的优先级，即基于代数表示的条件属性 c 是冗余属性还是非冗余属性。

$NewSig(c)$考虑的是基于信息表示的条件属性的重要度及冗余属性的权重不为 0，使权重的赋值更符合实际要求，更为合理。但文献[1]的这种权重确定方法却导致原为冗余属性的指标重要度高于原为非冗余属性的指标重要度，与基于代数表示的权重确定相矛盾。将 $NewSig(c)$和 $\mu(c)$结合起来引入权重定义，一方面考虑属性重要程度及各属性权重不为 0，另一方面，考虑原为非冗余属性的指标权重高于原为冗余属性的权重，使权重赋值更为合理。

定理 1 在决策表 $S=(U,C,D,V,f)$中，c1，c2C，若有 $Sig(c1)>0$，$Sig(c2)=0$，则 $NewW(c1)>NewW(c2)$。

证明：在决策表 $S=(U,C,D,V,f)$中，$\forall c1\in C$，有 $Sig(c1)>0$，说明条件属性 c1 是非冗余属性，列入高级优先级队列中，必 $\exists b\in C$，且 $Sig(b)=0$，使 $NewSig(b)$的值在冗余属性集中（低级优先级队列）为最大值，此时对 $\forall c2\in C$，且 $Sig(c2)=0$，皆有 $NewSig(b)\geqslant NewSig(c2)$，可得 $NewSig(b)\geqslant NewSig(c2)$，由定义 6 即可得 c1 的优先度 $\mu(c1)=NewSig(b)$，由以上两点可知，$\mu(c1)\geqslant NewSig(c2)$。另一方面，对 $c2\in C$，$Sig(c2)=0$，说明条件属性 c2 是非冗余属性，列入低级优先级队列中，由定义 6 可得 c2 的优先度 $\mu(c2)=0$；由定义 5 可知 $NewSig(c1)>0$，那么，

NewSig(c1)+μ(c1)>NewSig(c2)+μ(c2)成立，即 NewW(c1)>NewW(c2)成立。证毕。

定理1说明原为非冗余属性的指标权重比原为冗余属性的指标权重大，其重要度更高些。

定理1的逆命题不一定成立。

采用定义6和定义7计算表1各条件属性的新的权重如下：

$$\mu(c1)=\mu(c2)=\frac{26}{100};\ \mu(c3)=\mu(c4)=0;$$

NewW(c1)=0.31；NewW(c2)=0.40；NewW(c3)=0.19；NewW(c4)=0.10。

由以上计算结果可以看出，原为冗余属性的c3,c4的权重都不为0，原为非冗余属性的c1和c2的权重明显高于原为冗余属性的指标，得到的指标权重顺序是：NewW(c2)>NewW(c1)>NewW(c3)>NewW(c4)，与定义1和定义2计算的基于代数表示的粗糙集权重大小顺序保持一致。采用改进的条件信息熵权重确定方法，既能与基于代数表示的粗糙集重要度保持一致，又能避免属性权重为0的这种不合理的权重定义。

3 算例与分析

为了验证本文方法的有效性，本算例将对某市有代表性的10个行政机关进行政府信息公开公众满意度调查，对政府信息公开公众满意度评价进行计算，评价指标选自文献[12]中政府信息公开满意度评价指标集。设政府部门集合U={x1,x2,…,x10}，条件指标集合C={公平性、及时性、规范性、便利性}，简记为C={c1、c2、c3、c4}，决策指标集合D={公众满意}，简记为D={d}。依据这5个评价指标，公众分别对10个行政机关的政府信息公开公众满意度的评价指标打分，10代表非常满意，0代表非常不满意，公众满意指标值为1表示满意，0表示不满意，经过初步处理后得到原始评价信息表见表2，对表2的数据进行离散化处理，形成如表1的决策表S1=(U,C,D,V,f)。

表1 决策表 S1=(U,C,D,V,f)

	c1	c2	c3	c4	D
x1	1	2	2	3	1
x2	2	3	2	1	0
x3	1	2	2	3	0
x4	3	3	1	2	1
x5	1	3	2	1	0
x6	3	1	1	2	0
x7	3	3	2	2	1
x8	2	3	2	1	1
x9	2	3	2	1	0
x10	3	3	3	2	1

表2 原始信息表

	c1	c2	c3	c4	D
x1	5	6	7	8	1
x2	6	8	6	5	0
x3	5	6	6	8	0
x4	8	9	5	7	1
x5	5	8	7	5	0
x6	8	5	5	6	0
x7	8	9	7	7	1
x8	7	9	7	5	1
x9	6	8	6	5	0
x10	8	8	9	7	1

本算例将本文的权重确定方法和采用代数表示的粗糙集权重确定方法[11]、粗糙集条件信息熵的权重确定方法[1]确定的各条件属性（指标）的权重系数（见本文第3节），用线性加权法计算各对象的最终评价结果，见表3。

从代数表示粗糙集权重确定方法来看，政府部门x4,x7,x8,x10的评价较好，但从原始信息表2来看，x4的规范性和x8的便利性指标分值最低，最终评价值中没有体现，因为这两个指标的权重系数为0，是作为冗余属性处理的，显然不太合理。从粗糙集条件信息熵权重确定方法来看，这4个政府部门政府信息公开的评价值依然较高，其中x4、x8的评价值降低，因为粗糙集条件信息熵权重确定方法让每个指标的权重都不为0，但及时性、公平性指标表现最好的x7，其规范性、便利性指标的分值也为次优，x7的评价值却不是最高的，排在第二位。采用本文方法得到的评价值中x7的分值最高，x4的规范性、x8的便利性指标得分最低在最终评价值中也得到了体现，说明本文方法较符合实际结果，也表明了本文方法的有效性。

表3 三种权重确定方法的评价结果

对象 / 评价结果 / 方法	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
代数表示粗糙集权重	5.67	6.34	5.67	8.67	7.01	5.99	8.67	8.34	7.34	8
粗糙集条件信息熵权重	6.46	6.49	6.15	7.23	6.61	5.74	7.85	7.32	6.49	8.14
本文方法	6.08	6.7	5.89	7.73	6.58	6.03	8.11	7.6	6.7	8.09

4 结论

基于以上分析与算例，本文提出的方法比文献1的权重方法更优；从理论上讲本文的方法包含了原方法，并修正了原方法出现的冗余属性重要度过高的问题，而且也可以证明，本文的方法符合实际结果和具有有效性。

参考文献：

[1]鲍新中,张建斌,刘澄.基于粗糙集条件信息熵的权重确定方法[J].中国管理科学,2009,(17).

[2]Pawlak Z. Rough Sets[J].International Journal of Computer and Information Science,1982,11(5).

[3]黄健,张元标.粗糙集理论在科研能力综合评价中的应用[J].科技管理研究,2009,(2).

[4]史文利,高天宝,王树恩等.基于粗糙集理论的物流客户满意度评价研究[J].软科学,2008,(22).

[5]张金隆,卢新元等.基于粗糙集的投标风险分析与规避决策模型研究[J].管理学报,2008,5(1).

[6]高俊山,谷冬元,徐章艳等.一个Pawlak粗糙集冲突分析模型的改进[J].中国管理科学，2008,16(2).

[7]刘勇,熊蓉,褚健.Hash快速属性约简算法[J].计算机学报,2009,32(8).

[8]王熊彬,郑雪峰,徐章艳等.基于系统熵的属性约简的简化差别矩阵方法[J].计算机应用研究,2009,7(2).

[9]Neil Mac Parthala, Qiang Shen, Richard Jensen. A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction[J].IEEE Transactions on Knowledge and Data Engineering,2010,(22).

[10]Wang J,Wang J. Reduction Algorithms Based on Discernibility Matrix: The Ordered Attributes Method[J].Jounary Computer Science Technology,2001,16(6).

[11]曹秀英,梁静国.基于粗糙集理论的属性权重确定方法[J].中国管理科学,2002,10(5).

[12]段尧清,冯骞.政府信息公开满意度研究(Ⅱ)—基于结构方程的满意度模型构建[J].图书情报工作,2009,(53).

（责任编辑/亦 民）