一种基于粗糙集条件信息熵的多指标综合评价方法研究*

毛太田　肖　锎　邹　凯

内容提要: 针对多指标综合评价中各评价指标权重获取的问题,构建了基于粗糙集条件信息熵的多指标综合评价智能模型。考虑到实际综合评价应用分析中常见的小样本数据特点,将决策表进行合理分块,以层次式计算的方式基于粗糙集方法逐步实现对各个评价指标权重的客观求取。通过算例分析和实验比较,验证了该粗糙集智能评价模型的有效性与优越性。

关键词: 粗糙集; 综合评价; 信息熵; 指标权重
中图分类号: C829.22　　**文献标识码:** A　　**文章编号:** 1002-4565(2014)07-0092-05

A Research on Multiple-indicators Comprehensive Evaluation Method Based on Rough Set and Conditional Information Entropy

Mao Taitian　Xiao Kai　Zou Kai

Abstract: In order to get the reasonable and objective evaluation index weights, rough set theory based intelligent evaluation model is constructed. According to the problem of index-weight acquirement in accordance with the condition of small sample data, we try to divide the decision table into several blocks and use hierarchical way to get index weight values, based on rough set theory. At the end of this paper, experimental results prove the superiority of the rough set intelligent evaluation model.

Key words: Rough Set; Comprehensive Evaluation; Information Entropy; Index Weight

一、引言

粗糙集理论是 Z. Pawlak(1982)[1] 提出的一种数据推理方法,其具有不依赖样本数据之外的先验知识而真实反映数据本身所隐藏的信息、得到问题内在规律的独特优势[2]。本文拟将信息熵观点下的粗糙集计算方法引入到综合评价中,尝试构建一种智能化定量获取指标权重的层次式计算新算法,为综合评价提供更为科学有效的评价模型算法。

二、综合评价智能模型的构建

(一)粗糙集基本理论

定义 1[3]:称 $S=(U,A,V,f)$ 是一个决策表,其中,$U=\{u_1,u_2,\cdots,u_n\}$ 是一个非空、有穷、被称为全域的个体的集合;A 是非空、有穷的属性集合,$C\cup D=A$,C 为条件属性集,D 为决策属性集,$f: U\times A\rightarrow V$ 称为映射函数,对于属性 $a\in A$,有 $a: U\rightarrow V_a$,V_a 为属性 a 的值集,集合 $V=\bigcup_{a\in A}V_a$ 为属性集 A 的值域。

定义 2[4]:给定决策表 S,若有 $B\subseteq A$,则定义属性集 B 上的不可分辨关系 $IND(B)$ 为:

$$IND(B)=\{(u_i,u_j)\in U^2 \mid \forall b\in B, f(u_i,b)=f(u_j,b)\}.$$

定义 3[5]:在决策表 S 中,$\forall X\subseteq U$ 且 $X\neq\varnothing$,则定义集合 X 在属性集 $B\subseteq A$ 上的下近似划分集 $\underline{B}(X)$ 和上近似划分集 $\overline{B}(X)$ 为:

$$\underline{B}(X)=\{u_i\in U \mid [u_i]_B\subseteq X\}$$

$$\overline{B}(X)=\{u_i\in U \mid [u_i]_B\cap X\neq\varnothing\}$$

其中,$[u_i]_B=\{u_j \mid (u_i,u_j)\in IND(B)\}$,$\underline{B}(X)$ 也称为 X 的 B 正域,记作:$Pos_B(X)$。

* 本文获国家自然科学基金项目"面向复杂类型数据的粒计算方法、模型及其多属性群决策分析"(71101096)资助。

定义 4[4]: 在决策表 S 中，若有 $B \subseteq C, Y \in U/IND(D)$，则决策属性 D 的 B 正域 $Pos_B(D)$ 定义为:

$$Pos_B(D) = \bigcup_{Y \in U/IND(D)} \underline{B}(Y)$$

定义 5[3]: 在决策表 S 中，若 $U/IND(C) = \{X_1, X_2, \cdots, X_q\}$，$U/IND(D) = \{Y_1, Y_2, \cdots, Y_p\}$，则对象集 U 在条件属性集 C 下相对于决策属性 D 划分的信息熵定义为:

$$I(D \mid C) = -\sum_{s=1}^{q} \frac{Card(X_s)}{Card(U)} \times \sum_{k=1}^{p} \frac{Card(Y_k \cap X_s)}{Card(X_s)} \times \log_{10}\left(\frac{Card(Y_k \cap X_s)}{Card(X_s)}\right)$$

其中，Card(*) 表示集合的基数。

定义 6[3]: 在决策表 S 中，定义属性 $c \in C$ 粗糙集信息熵意义下的重要度为:

$$SGF(c) = I(D \mid C) - I(D \mid C - \{c\})$$

(二) 综合评价决策分析算法

1. 数据预处理。在综合评价指标体系中，有些指标是可以量化的，有些指标是不能够量化的。对于不能量化的指标，可以采取专家打分的形式确定; 对于可以量化的指标，则应根据实际值的大小予以评分。由于不同的指标在数量级和量纲上有差异，因此，评分法应消除属性间的不可公度性，以保证决策表中数据关系的一致性。

本文以综合评价指标体系中只存在一级指标和二级指标为研究对象。把条件属性集(指标集) 描述为 $C = \{C_1, C_2, \cdots, C_z\}$，对于 $C_x(x = 1, 2, \cdots, z)$ 作为一级指标条件属性，其中包含若干二级指标条件属性，则可以进一步描述为 $C_x = \{c_{x1}, c_{x2}, \cdots, c_{xv}\}$。由此，$\forall u_i \in U$，则评价对象在二级指标条件属性 c_{xv} 下的评分取值可以描述为 l_{xv}^i。对于评价对象 u_i 在条件属性 c_{xv} 下的评分值 l_{xv}^i，我们拟进行如下的一致化处理[7]。

(1) 如果评价指标属性为成本型属性，则对属性值 l_{xv}^i 的一致化处理描述为:

$$\widetilde{l_{xv}^i} = 100 \times [\max(l_{xv}) - l_{xv}^i] / [\max(l_{xv}) - \min(l_{xv})]$$

(2) 如果评价指标属性为效益型属性，则对属性值 l_{xv}^i 的一致化处理描述为:

$$\widetilde{l_{xv}^i} = 100 \times (l_{xv}^i - \min(l_{xv})) / (\max(l_{xv}) - \min(l_{xv}))$$

(3) 如果评价指标属性为特定最优值取值属性，则对属性值 l_{xv}^i 的一致化处理描述为:

$$\widetilde{l_{xv}^i} = \begin{cases} 100 \times [1 - (\overline{l_{xv}^i} - l_{xv}^i) / (\overline{l_{xv}^i} - l'^i_{xv})] & \overline{l_{xv}^i} > l_{xv}^i > l'^i_{xv} \\ 100, \overline{l_{xv}^i} = l_{xv}^i & \\ 100 \times [1 - (l_{xv}^i - \overline{l_{xv}^i}) / (l''^i_{xv} - \overline{l_{xv}^i})] & \overline{l_{xv}^i} < l_{xv}^i < l''^i_{xv} \\ 0, \text{其他} & \end{cases}$$

其中，$\max(l_{xv})$ 为原始数据集中属性 c_{xv} 下的最大数据取值; $\min(l_{xv})$ 为原始数据集中属性 c_{xv} 下的最小数据取值; $\overline{l_{xv}^i}$ 为属性 c_{xv} 下的最优值; l'^i_{xv} 为属性 c_{xv} 下无法接受下限; l''^i_{xv} 为属性 c_{xv} 下无法接受上限。

2. 连续型数据离散化。由于粗糙集只能处理离散化的数据，因此，需要对连续型数据进行离散化处理，本文采用等距离法[8]，具体步骤如下。

(1) 条件属性 c_{xv} 在离散化时的取值区间长度计算为:

$$l_{xv}^* = \frac{\max(\widetilde{l_{xv}^i}) - \min(\widetilde{l_{xv}^i})}{m} \quad (1)$$

其中，l_{xv}^* 为区间的长度; $\max(\widetilde{l_{xv}^i})$ 为属性 c_{xv} 中的最大评分值; $\min(\widetilde{l_{xv}^i})$ 为属性 c_{xv} 中的最小评分值; m 为设定的离散化区间数目。

(2) 对于对象 u_i，在条件属性 c_{xv} 下的离散化结果计算为:

$$t_{xv}^i = < \frac{\widetilde{l_{xv}^i} - \min(\widetilde{l_{xv}^i})}{l_{xv}^*} > \quad (2)$$

其中 t_{xv}^i 为对于对象 u_i，在条件属性 c_{xv} 下取值的离散化结果 "< * >" 表示向上取整。

3. 构建决策表。把数据离散化的结果转化成为决策表 $S = (U, A, V, f)$，$U = \{u_1, u_2, \cdots, u_n\}$ 表示各个评价对象的集合，条件属性评价指标集 $C = \{C_1, C_2, \cdots, C_z\}$，其中 $C_x(x = 1, 2, \cdots, z)$ 为一级指标条件属性，该一级指标条件属性包含若干二级指标条件属性，可以描述为 $C_x = \{c_{x1}, c_{x2}, \cdots, c_{xv}\}$，决策属性集 $D = \{d_1, d_2, \cdots, d_r\}$。

4. 客观权重确定。通过把基于信息熵观点下的粗糙集计算方法引入综合评价智能分析中，获取属性(也即指标) 的客观权重。考虑到综合评价实际应用分析中常见的评价对象数据规模通常较少，而相应的评价分析指标数目较多，因此在利用粗糙集方法客观计算评价指标的权重

时，考虑基于一级指标集进行分块划分，以层次式计算的方式逐步实现对各个二级指标权重的确定，为小样本数据问题粗糙集求解提供了一个全新的方案。

在信息熵观点下，粗糙集运算直观性较强，不同属性的重要性可以基于信息熵的运算得到相应的定量化数值。即通过从决策表中剔除某条件属性，再考察在该属性缺失的情况下整个决策分类信息熵的变化情况。如果剔除后变化较大，则说明该属性的重要性大；反之，重要性小。

算法 1：局部分块下的各二级指标条件属性权重值计算。

输入：决策表 $S = (U, A, V, f)$。

输出：各一级指标下的二级指标条件属性相对权重值。

步骤：

①依次选取决策表中的一级指标 $C_x(x = 1, 2, \cdots, z)$，求取对象集 U 在一级指标 C_x 分块下的划分结果 $U/IND(C_x)$ 和在决策属性 D 上的划分结果 $U/IND(D)$，计算信息熵 $I(D \mid C_x)$；

②在选定的一级指标 C_x 下，依次剔除该一级指标下的二级指标条件属性 $c_{xh}(h = 1, 2, \cdots, v)$，求取划分结果 $U/IND(C_x - \{c_{xh}\})$，并计算信息熵 $I(D \mid C_x - \{c_{xh}\})$；

③计算一级指标 C_x 分块下的各二级指标条件属性 c_{xv} 在粗糙集意义下的相对重要度：$SGF(\{c_{xh}\}) = I(D \mid C_x) - I(D \mid C_x - \{c_{xh}\})$；

④计算在一级指标 C_x 分块下的各个二级指标条件属性 c_{xh} 的相对权重值：

$$\omega(c_{xh}) = \frac{SGF(\{c_{xh}\})}{\sum_{h=1}^{v} SGF(\{c_{xh}\})}$$

算法 2：各一级指标的权重值计算。

输入：决策表 $S = (U, A, V, f)$。

输出：各一级指标的权重值。

步骤：

①求取对象集 U 在条件属性集 C 上的划分结果 $U/IND(C)$ 和在决策属性 D 上的划分结果 $U/IND(D)$，以及在每次剔除条件属性集 C 下的一级指标 C_x 后的划分结果 $U/IND(C - C_x)$；

②计算条件属性集 C 上的划分相对决策属性 D 上的划分的信息熵 $I(D \mid C)$，以及在依次剔除一级指标 C_x 后，计算条件属性集 $C - C_x$ 上的划分相对决策属性 D 上的划分的信息熵 $I(D \mid C - C_x)$；

③计算各一级指标 C_x 的重要度：

$SGF(C_x) = I(D \mid C) - I(D \mid C - C_x)$；

④计算各一级指标 C_x 的权重值：

$$\omega(C_x) = \frac{SGF(C_x)}{\sum_{x=1}^{z} SGF(C_x)}$$

算法 3：待评价对象的综合评价值。

输入：决策表 $S = (U, A, V, f)$，局部分块下的各二级指标条件属性相对权重值 $\omega(c_{xv})$ 和一级指标权重值 $\omega(C_x)$。

输出：待评价对象 $u_i \in U$ 的综合评价值 K_i。

步骤：

①计算全局意义下的各二级指标条件属性的最终权重值 $\bar{\omega}(c_{xh}) = \omega(C_x) \times \omega(c_{xh})$；

②计算各个待评价对象 K_i 的综合评价结果值

$$K_i = \sum_{x=1}^{z} \sum_{h=1}^{v} \bar{\omega}(c_{xh}) \times \widetilde{l_{xh}^{i}}。$$

三、实例分析

为了验证基于粗糙集条件信息熵的多指标综合评价智能模型的合理有效性，本文对湖南省 10 所地方高校的债务化解能力进行评价，评价指标选自文献[3]。设 10 个地方高校为 $U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$，评价指标包括 2 个一级指标条件属性，“高校层面组织管理、政府层面组织管理”，记为 $C = \{C_1, C_2\}$；7 个二级指标条件属性，分别为：“高校层面组织管理”下的“成立校内化债工作领导小组、制定化债工作规划、建立校内化债工作责任制”，记为 $C_1 = \{c_{11}, c_{12}, c_{13}\}$；“政府层面组织管理”下的“成立政府层面化债工作领导小组、核定高校基本建设总体规模、制定高校化债工作目标责任制、建立高校化债控债工作机制”，记为 $C_2 = \{c_{21}, c_{22}, c_{23}, c_{24}\}$；将“专家满意度”作为决策属性 D，记为 $D = \{d_1, d_2, \cdots, d_i\}$。选取高校管理领域的 10 位专家根据自己的主观感受分别对 10 所地方高校的 7 个二级指标进行打分（100 分表示非常满意，0 分表示非常不满意），将各个专家对每个二级指标的打分结果进行综合平均，得到每个二级指标的专家主观评价值 l_{xh}^{i}；根据高校 u_i 的二级指标专家主观评价值定义该高校的专家满意度为：

$$d_i = \sum l_{xh}^i / e \quad (3)$$

其中，e 为二级指标的数目; d_i 为高校 u_i 的专家满意度。

通过公式(3)计算专家满意度，初始信息表见表1。由于粗糙集只能处理离散数据，因此将二级指标和决策属性均分为3个等级，利用式(1)确定各个指标的离散化区间，并根据式(2)确定各个指标的离散化结果，决策表的离散化结果见表2。

表1　　初始信息表

检 / 指标	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}
C_{11}	91	70	90	80	100	85	60	85	80	90
C_{12}	75	81	60	100	75	85	85	90	80	60
C_{13}	83	89	77	65	85	100	62	85	76	77
C_{21}	80	25	55	100	50	25	55	25	75	25
C_{22}	68	80	100	75	89	88	75	85	89	85
C_{23}	85	89	65	65	100	88	85	100	76	100
C_{24}	70	90	80	75	100	90	72	90	89	90
D	78.86	74.86	75.29	80	85.57	80.14	70.57	80	80.71	75.29

表2　　数据离散化后的决策表

高校	条件属性集 C							决策属性 D
	C_1			C_2				
	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{23}	C_{24}	
u_1	3	2	2	3	1	2	1	2
u_2	1	2	3	1	2	3	3	1
u_3	3	1	2	2	3	1	2	1
u_4	2	3	1	3	1	1	1	2
u_5	3	2	2	1	2	3	3	3
u_6	2	2	3	1	2	2	3	2
u_7	1	2	1	2	1	2	1	1
u_8	2	3	2	1	2	3	3	2
u_9	2	2	2	2	2	1	2	3
u_{10}	3	1	2	1	2	3	3	1

本文分别以“高校层面组织管理”和“政府层面组织管理”将决策表进行分块，并基于信息熵观点下的粗糙集方法分别求取“高校层面组织管理”与“政府层面组织管理”下的局部分块二级指标权重。由表2，根据算法1可计算求得“高校层面组织管理”一级指标下的各个二级指标的相对权重值为:

$U/IND(C_1) = \{\{1,5\},\{2\},\{3,10\},\{4\},\{6\},\{7\},\{8\},\{9\}\}$

$U/IND(D) = \{\{1,4,6,8\},\{2,3,7,10\},\{5,9\}\}$

$U/IND(C_1 - \{c_{11}\}) = \{\{1,5,9\},\{2,6\},\{3,10\},\{4\},\{7\},\{8\}\}$

$U/IND(C_1 - \{c_{12}\}) = \{\{1,3,5,10\},\{2\},\{4\},\{6\},\{7\},\{8,9\}\}$

$U/IND(C_1 - \{c_{13}\}) = \{\{1,5\},\{2,7\},\{3,10\},\{4,8\},\{6,9\}\}$

$$I(D \mid C_1) = \frac{1}{5}\ln^2$$

$$I(D \mid C_1 - \{c_{11}\}) = \frac{3}{10}\ln^3$$

$$I(D \mid C_1 - \{c_{12}\}) = \frac{4}{5}\ln^2$$

$$I(D \mid C_1 - \{c_{13}\}) = \frac{2}{5}\ln^2$$

$$SGF(\{c_{11}\}) = I(D \mid C_1) - I(D \mid C_1 - \{c_{11}\}) = -\frac{3}{10}\ln^3 + \frac{1}{5}\ln^2$$

$$SGF(\{c_{12}\}) = -\frac{3}{5}\ln^2$$

$$SGF(\{c_{13}\}) = -\frac{1}{5}\ln^2$$

$\omega(c_{11}) = 0.256$; $\omega(c_{12}) = 0.558$; $\omega(c_{13}) = 0.186$

同理可得，“政府层面组织管理”一级指标下的各个二级指标的相对权重值为:

$\omega(c_{21}) = 0.356$，$\omega(c_{22}) = 0.356$

$\omega(c_{23}) = 0.287$，$\omega(c_{24}) = 0$

根据以上计算结果分析，地方高校债务化解能力指标体系下的“建立高校化债控债工作机制”为冗余属性，应去除该冗余属性。

然后，在整个决策表中分别剔除一级指标“高校层面组织管理”与“政府层面组织管理”下的所有二级指标，并基于信息熵观点下的粗糙集方法求取各一级指标的权重值。根据算法2可得:

$U/IND(C) = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\}\}$

$U/IND(C - C_1) = \{\{1\},\{2,5,8,10\},\{3\},\{4\},\{6\},\{7\},\{9\}\}$

$U/IND(C - C_2) = \{\{1,5\},\{2\},\{3,10\},\{4\},\{6\},\{7\},\{8\},\{9\}\}$

$$I(D \mid C) = 0$$

$$I(D \mid C - C_1) = \frac{3}{5}\ln^2$$

$$I(D \mid C - C_2) = \frac{1}{5}\ln^2$$

$$SGF(C_1) = -\frac{3}{5}\ln^2$$

$$SGF(C_2) = -\frac{1}{5}\ln^2$$

$\omega(C_1) = 0.75$;$\omega(C_2) = 0.25$

最后把一级指标下获得的局部分块下的二级指标权重与一级指标权重进行融合，得到全局意义下的各个二级指标最终客观权重值，根据算法 3 可得:

$\overline{\omega}(c_{11}) = 0.192$; $\overline{\omega}(c_{12}) = 0.419$;

$\overline{\omega}(c_{13}) = 0.140$; $\overline{\omega}(c_{21}) = 0.089$;

$\overline{\omega}(c_{22}) = 0.089$; $\overline{\omega}(c_{23}) = 0.072$

最终得到 10 所地方高校债务化解能力综合评价值分别为:

$K_1 = 79.809$,$K_2 = 75.592$

$K_3 = 71.675$,$K_4 = 86.615$

$K_5 = 82.096$,$K_6 = 82.328$

$K_7 = 73.505$,$K_8 = 82.920$

$K_9 = 79.588$,$K_{10} = 70.190$

四、方法比较

为了进一步验证本文方法的合理性和优越性，下面将本文方法与周志远和沈固朝(2012)[9]中所提出的方法进行对比分析。采用他们的方法基于本文的评价指标与数据对 10 所地方高校的债务化解能力进行评价。为了验证本文模型的合理有效性，特定义评价决策方案区分度为:

$$\eta = \sum_{i=1}^{n}(K_i - \overline{K})^2, \overline{K} = \sum_{i=1}^{n} K_i/n \qquad (4)$$

其中，n 为高校的数目。

区分度越大意味着待评价地方高校间的优势可分辨性越大，也直接表明该评价方法的优越性。通过式(4)计算上述两种评价方法下决策方案的区分度，结果显示，本文方法的评价决策方案区分度更大，表明更具优越性，也表明本文模型的合理有效。

五、结束语

本文把信息熵观点下的粗糙集方法引入综合评价分析中，在已有的研究基础上建立了基于粗糙集条件信息熵的智能评价模型，客观获取各评价指标的权重值，为后续的综合评价分析奠定基础。实例表明，该粗糙集智能评价模型是切实可行且有效的，为综合评价提供了新的智能化评价方法。

参考文献

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

[2] Yee Leung, Manfred M. Fischer, Wei-Zhi Wu, Ju-Sheng Mi. A rough set approach for the discovery of classification rules in interval-valued information systems [J]. International Journal of Approximate Reasoning, 2008,5(2):233 - 246.

[3] 郑学敏. 一种基于粗糙集理论的多指标综合评价方法 [J]. 统计与决策,2010(5):37 - 39.

[4] 朱红灿,陈能华. 粗糙集条件信息熵权重确定方法的改进 [J]. 统计与决策,2011(8):154 - 156.

[5] 谭旭,唐云岚,陈英武. 基于粗糙集的区间型数据离散化算法 [J]. 系统工程理论与实践,2009,29(6):157 - 165.

[6] 高维春,谭旭. 决策属性未知下的学生评教粗糙集分析 [J]. 计算机工程与应用,2012,48(9):238 - 241.

[7] 岳超源. 决策理论与方法 [M]. 北京:科学出版社,2006.

[8] 孙红丽,李永宁. 地方高校债务化解绩效评价指标体系的构建 [J]. 统计与决策,2012(14):53 - 55.

[9] 周志远,沈固朝. 粗糙集理论在情报分析指标权重确定中的应用 [J]. 情报理论与实践,2012,35(9):61 - 65.

作者简介

毛太田，男，1971 年生，湖南永州人，2008 年毕业于国防科学技术大学，获管理学博士学位，现为湘潭大学公共管理学院副教授，统计学专业硕士生导师。研究方向为信息分析与评价、统计信息管理。

肖锎，男，1988 年生，湖南娄底人，湘潭大学公共管理学院硕士研究生。研究方向为情报理论与智能决策评价、统计信息管理。

邹凯，男，1965 年生，湖南新化人，2008 年毕业于国防科学技术大学，获管理学博士学位，现为湘潭大学公共管理学院教授、副院长，统计学专业博士生导师。研究方向为统计信息管理、管理信息系统。

(责任编辑: 曹 麦)