

基于粗集优势关系的属性赋权相对熵优化模型

毛军军^{1a,1b}, 李侠^{1a}, 吴涛^{1a,2}

(1. 安徽大学 a. 数学科学学院; b. 计算智能与信号处理教育部重点实验室, 合肥 230039;

2. 南京大学计算机软件新技术国家重点实验室, 南京 210093)

摘要: 针对传统粗集理论中属性赋权不一致, 甚至相悖的问题, 把2个概率分布的相对熵扩展到任意2个单维向量的相对熵, 并将相对熵视作一种距离。通过定义属性重要度的代数观和粒度观确定优化权重的取值范围, 根据各方案的属性值尽可能靠近理想值、远离负理想值的原理, 建立单目标赋权优化模型。针对等价关系的局限性, 将优势关系引入属性权重确定方法中。基于优势关系的序信息系统, 将代数观下和粒度观下的权重通过相对熵优化模型进行耦合, 得到多属性决策中属性权重的优化解。算例分析结果证明了该模型的有效性。

关键词: 粗糙集; 优势关系; 属性依赖度; 粒度; 相对熵

Optimal Attribute Weighting Relative Entropy Model Based on Dominance Relation in Rough Set

MAO Jun-jun^{1a,1b}, LI Xia^{1a}, WU Tao^{1a,2}

(1a. School of Mathematical Sciences; 1b. Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education,

Anhui University, Hefei 230039, China; 2. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

【Abstract】 Aiming at the difference and antinomy of the weighting in the rough set, the relative entropy of two probability distributions is extended to the relative entropy of two one-dimension vectors. The relative entropy is viewed as a distance measurement. The values range of optimal weights are determined by determinations of attribute importance in algebra view and knowledge granularity view respectively, and a single objective optimization model is established on the grounds that attribute values of alternatives are far away from negative ideal values and as close as to ideal values. On the other hand, since the limitation of equivalence relations, dominance relations are introduced to the method of determining the attribute weights. The weights in algebra view and knowledge granularity view are carried on the organic integration by the relative entropy optimization model based on dominance relations, in order to obtain optimal solution of the attributes in Multiple Attribute Decision Making(MADM). The analysis result indicates the validity and efficiency of the model.

【Key words】 rough set; dominance relation; attribute dependency; granularity; relative entropy

DOI: 10.3969/j.issn.1000-3428.2011.15.039

1 概述

粗糙集理论^[1]是一种新型软计算工具。由该理论知, 知识库中描述知识的属性不是同等重要的, 甚至有些是冗余的。所谓属性依赖度, 就是从条件属性集中删除某属性, 然后考察没有该属性的情况下决策的分类发生的变化的度量。经典粗糙集理论是基于等价关系建立的, 但等价关系的应用领域极其有限。要定义论域上的等价关系, 属性值必须是离散的, 但在实际问题中属性值大多是连续的, 其中部分具有偏序关系。利用经典粗糙集方法处理这类信息系统时, 需先将其离散化^[2], 导致信息丢失。于是人们将等价关系放宽为相容关系、相似关系^[3]等。文献[4]提出了基于优势关系的粗糙集研究方法(DRSA), 利用优势关系建立序信息系统有助于处理连续属性和偏序关系的问题。

在综合评判和决策分析中, 权重是各指标在评估决策中所起的作用大小的度量, 它直接影响综合评估和决策的最终结果。粗糙集理论出现后, 有些学者提出了根据该理论中属性重要性的概念来确定属性(指标)权重。但粗糙集理论中属性重要性的几个定义不具备一致性, 进而得到的属性权重也不具有一致性, 有时甚至是相悖的。由属性依赖度与知识粒度的相关性知, 这2种定义确定的权重具有互补的特性: 序信息系统中属性依赖度随着分辨能力的增强(属性集中属

性增加)而单调增, 知识的粒度随着分辨能力的增强而单调减^[5]。相对熵用来度量2个概率分布的符合程度^[6], 本文将两组权重看作概率分布, 根据相对熵的定义建立优化模型, 得到权重的优化解。

2 概念准备

在 Pawlak 近似空间意义下的信息系统是基于等价关系(二元不可区分关系)的, 若对每个属性值域都有按递增或递减的一个偏序关系, 这时就需建立基于优势关系的信息系统, 即序信息系统。

定义 1(序信息系统) 在一个信息系统中, 如果在某个属性值域上建立一个偏序关系, 称这个属性为一个准则。当所有的属性都为准则时, 该信息系统称为序信息系统^[1]。

设在信息系统 $S = (U, A, V, f)$ 中属性 u 是一个准则, 并且在 u 的值域上建立的偏序关系是“ \geq_u ”。于是对于对象 x, y ,

基金项目: 国家自然科学基金资助项目(60675031); 中国博士后科学基金资助项目(20070411028); 安徽省高等学校省级自然科学基金资助项目(KJ2008B093); 安徽大学学术创新团队基金资助项目(KJTD001B); 安徽省高等学校青年基金资助项目(2011SQRL186)

作者简介: 毛军军(1973—), 女, 副教授、博士, 主研方向: 智能计算, 粗糙集理论; 李侠, 硕士研究生; 吴涛, 教授、博士后

收稿日期: 2011-01-24 **E-mail:** maojunjun@ahu.edu.cn

说 $x \geq_u y$ 表示 x 至少和 y 关于准则 u 是一样好的, 或者说 x 优于 y 。

不失一般性, 取属性的值域为实数, 定义关系 “ \geq_u ” 为: $x \geq_u y \Leftrightarrow f(x, u) \geq f(y, u)$ 。

于是对于属性集 $B \subseteq A$, $x \geq_B y$ 是指 x 关于属性集 B 中的所有准则都优于 y 。

定义 2(优势关系) 设 $S = (U, A, V, f)$ 为序信息系统, 对于 $B \subseteq A$, 令 $R_B = \{(x, y) \in U \times U \mid f_m(x) \geq f_m(y), \forall a_m \in B\}$, 则 R_B 称为序信息系统 $S = (U, A, V, f)$ 的优势关系^[1]。若记:

$$[x_i]_B = \{x_j \in U \mid (x_i, x_j) \in R_B\} = \{(x, y) \in U \times U \mid f_m(x) \geq f_m(y), \forall a_m \in B\}$$

$$U/R_B = \{[x_i]_B \mid x_i \in U\}$$

则称 $[x_i]_B$ 为对象 x_i 的优势类; U/R_B 为对象集关于属性集 B 的一个分类。

定义 3(上近似、下近似) 对于任意 $X \subseteq U$, 定义 X 关于优势关系 R_B 的下近似和上近似分别为^[1]:

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

$$\overline{R}_B(X) = \{x_i \in U \mid [x_i]_B \cap X \neq \emptyset\}$$

同样地, 优势关系下的正域定义为:

$$Pos_B(X) = \underline{R}_B(X) = \{x_i \in U \mid [x_i]_B \subseteq X\}$$

表示根据知识 B 判断肯定属于 X 的 U 中的元素组成的集合。

3 序信息系统下代数观和粒度观的属性权重

3.1 基于代数观的属性权重

定义 4 令 P 和 Q 为 U 中的优势关系, Q 的 P 正域记为 $Pos_P(Q)$, 即:

$$Pos_P(Q) = \bigcup_{x \in U/Q} PX \tag{1}$$

Q 和 P 的依赖度定义为 $\gamma_P(Q) = \frac{|Pos_P(Q)|}{|U|}$, 其中, $0 \leq \gamma_P(Q) \leq 1$ 。

随着 P 分辨能力的增强, Q 的 P 正域的基数增大, 由依赖度的定义可得, 若 $P, P' \subseteq A$ 且 $P' \subseteq P$, 则有: $Pos_{P'}(Q) \subseteq Pos_P(Q)$ 。即属性依赖度随着分辨能力的增强(属性集中属性增加)而单调增大。

定义 5(优势关系下属性重要性的代数观) 设 $S = (U, A, V, f)$ 为序信息系统, $A = C \cup D$, 且 $C \cap D = \emptyset$, C 和 D 分别为条件属性集和决策属性集, 属性子集 $C' \subseteq C$ 关于 D 的重要性定义为:

$$\eta_D(C') = \gamma_C(D) - \gamma_{C-C'}(D) \tag{2}$$

特别地, 当 $C' = \{a\}$ 时, 属性 $a \in C$ 关于 D 的重要性为:

$$\eta_D(a) = \gamma_C(D) - \gamma_{C-\{a\}}(D)$$

在序信息系统 $S = (U, A, V, f)$ 中, 设条件属性集 $C = \{a_1, a_2, \dots, a_m\}$, 则属性 a_j 的权重为:

$$\alpha_j = \frac{\eta_D(a_j)}{\sum_{j=1}^m \eta_D(a_j)}$$

其中, $0 \leq \alpha_j \leq 1$; $\sum_{j=1}^m \alpha_j = 1$ 。

3.2 基于粒度观的属性权重

定义 6(序信息系统中知识的粒度) 设 $S = (U, A, V, f)$ 为序信息系统, 且 $B \subseteq A$, 则知识 B 的粒度定义为^[5]:

$$GK(B) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |[x_i]_B| \tag{3}$$

由文献[5]可知, 在序信息系统 $S = (U, A, V, f)$ 中, 若 $P, Q \subseteq A$ 且 $Q \subseteq P$, 则有: $GK(P) \leq GK(Q)$ 。由此可知, 优势关系下的序信息系统中知识的粒度随着分辨能力的增强而单调减小。

定义 7(优势关系下的属性重要性的粒度观) 设 $S = (U, A, V, f)$ 为序信息系统, $A = C \cup D$, 且 $C \cap D = \emptyset$, C 和 D 分别为条件属性集和决策属性集, 属性子集 $B \subseteq C$, 则对任意属性 $a \in C - B$ 的重要性 $Sig(a, B, D)$ 定义为:

$$Sig(a, B, D) = GK(B) - GK(B \cup \{a\}) \tag{4}$$

在序信息系统 $S = (U, A, V, f)$ 中, 设条件属性集 $C = \{a_1, a_2, \dots, a_m\}$, 则属性 a_j 的权重为:

$$\beta_j = \frac{Sig(a_j, C - \{a_j\}, D)}{\sum_{j=1}^m Sig(a_j, C - \{a_j\}, D)}$$

其中, $0 \leq \beta_j \leq 1$, $\sum_{j=1}^m \beta_j = 1$ 。

根据优势关系下属性重要性的代数定义和粒度定义, 得到 2 种属性权重的定义, 而这 2 种定义具有一定的互补性: 属性依赖度随着分辨能力的增强(条件属性集中属性增加)而单调增大, 知识的粒度随着分辨能力的增强而单调减小。为综合考虑 2 种定义的特征, 确定权重的优化解, 本文建立一种相对熵属性赋权优化模型。

4 属性赋权相对熵优化模型

相对熵是 2 个随机分布之间的距离的度量^[6], 也称为交叉熵或 Kullback-Leibler 距离。

定义 8 离散概率分布 $p(x)$ 和 $q(x)$ 之间的相对熵定义为^[6]:

$$D(p \parallel q) = \sum p(x) \ln \frac{p(x)}{q(x)} = E_p \ln \frac{p(X)}{q(X)}$$

当且仅当 $p = q$ 时, $D(p \parallel q) = 0$ 。规定 $0 \ln \frac{0}{q} = 0$ 和

$p \ln \frac{p}{0} = \infty$ (基于连续性假设), 且相对熵满足非负性要求。

由定义 8 可见, $D(p \parallel q) \neq D(q \parallel p)$, 即相对熵不对称, 因此, 它不是真正的距离度量, 但将相对熵视作分布间的“距离”往往很有用。

类似定义两向量之间的相对熵:

定义 9 同维向量 $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_l\}$ 和 $\beta = \{\beta_1, \beta_2, \dots, \beta_l\}$ 的相对熵定义为:

$$D(\alpha \parallel \beta) = \sum_{i=1}^l \alpha_i \ln \frac{\alpha_i}{\beta_i}$$

当且仅当 $\alpha_i = \beta_i (i = 1, 2, \dots, l)$ 时, $D(\alpha \parallel \beta) = 0$ 。并且规定 $0 \ln \frac{0}{\beta_i} = 0$ 和 $\alpha_i \ln \frac{\alpha_i}{0} = \infty$ 。由于 $\sum_{i=1}^l \alpha_i$ 和 $\sum_{i=1}^l \beta_i$ 不一定满足值为

1, 即 $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_l\}$ 和 $\beta = \{\beta_1, \beta_2, \dots, \beta_l\}$ 可能不能视为概率分布, 此时相对熵不满足非负性。为满足非负性要求, 笔者加一个序条件: $\alpha_i \geq \beta_i (i = 1, 2, \dots, l)$, 由定义 9 可知, 只要满足 $\alpha_i \geq \beta_i (i = 1, 2, \dots, l)$, 就有 $D(\alpha \parallel \beta) \geq 0$ 。

对于序信息系统 $S = (U, A, V, f)$, 设其方案集 $U = \{x_1, x_2, \dots, x_n\}$, 第 i 个方案 x_i 对第 j 个属性 a_j 的属性值记为 c_{ij} , 矩阵 $A = (c_{ij})_{n \times m}$ 称为属性矩阵或决策矩阵。属性通常可分为效益型、成本型、固定型、区间型。由于不同的属性一般具有不同的量纲, 在决策之前首先要将属性指标做无量纲化处理。记无量纲化处理后的矩阵为 $D' = (d_{ij})_{n \times m}$, 称为规范化的决策矩阵, d_{ij} 表示第 i 个方案 x_i 对第 j 个属性 a_j 的规范

化属性值, 显然 $d_{ij} \in [0, 1]$ 且越大越好。令 $D_j = \max_{1 \leq i \leq n} \{d_{ij}\}$, $d_j = \min_{1 \leq i \leq n} \{d_{ij}\}$, 则称向量 $D = \{D_1, D_2, \dots, D_m\}$ 和 $d = \{d_1, d_2, \dots, d_m\}$ 分别为 m 个属性的理想值和负理想值。

根据定义 9, 在确定属性权重的一个基本思想就是使每个方案的属性值更靠近理想值, 而远离负理想值, 即每个方案的属性值与理想值的相对熵尽可能小, 与负理想值的相对熵尽可能大, 则有如下多目标规划问题:

$$\begin{cases} \min F(W) = \sum_{i=1}^n \sum_{j=1}^m w_j D_j \ln \frac{D_j}{d_{ij}} \\ \max F(W) = \sum_{i=1}^n \sum_{j=1}^m w_j D_j \ln \frac{d_{ij}}{d_j} \end{cases} \quad (5)$$

$$\text{s.t.} \begin{cases} \sum_{j=1}^m w_j = 1 \\ 0 \leq w_j^- \leq w_j \leq w_j^+ \leq 1 \end{cases}$$

其中, $W = \{w_1, w_2, \dots, w_m\}$; w_j^- 和 w_j^+ 分别为第 j 个属性权重 w_j 的上界和下界。

由于 $D_j \geq d_{ij}$, $d_{ij} \geq d_j$ ($i=1, 2, \dots, n$; $j=1, 2, \dots, m$), 因此 $F(W)$ 满足非负性要求。

对于该多目标规划模型, 可将其转化为如下单目标规划问题:

$$\min H(W) = \sum_{i=1}^n \sum_{j=1}^m w_j D_j \ln \frac{d_{ij}}{D_j} / \sum_{i=1}^n \sum_{j=1}^m w_j d_{ij} \ln \frac{d_{ij}}{d_j} \quad (6)$$

$$\text{s.t.} \begin{cases} \sum_{j=1}^m w_j = 1 \\ 0 \leq w_j^- \leq w_j \leq w_j^+ \leq 1 \end{cases}$$

由文献[7]可知, 模型(6)的最优解一定是模型(5)的有效解, 故可以通过求解模型(6)的最优解获得模型(5)的有效解, 即为多属性决策问题的权重系数。

5 算例分析

数据来自《中国统计年鉴 2009》中北京、上海、江苏、浙江、福建、安徽、和四川 7 个地区的邮电业务量。其中, a_1 表示固定电话长途通话长(亿分钟); a_2 表示移动电话长途通话时长(亿分钟); a_3 表示 IP 电话通话时长(亿分钟); a_4 表示移动短信业务量(亿条); a_5 表示互联网上网人数(万人); a_6 表示移动电话年末用户(万户); a_7 表示固定电话年末用户(万户)。限于篇幅, 具体数据在此不罗列。

优势关系下约简后, 得到的属性集为 $\{a_3, a_4, a_5, a_7\}$ 。由定义 2 和式(1)、式(2)可得: $\gamma_C(C) = 1, \gamma_{C-\{a_3\}}(C) = \frac{5}{7}, \gamma_{C-\{a_4\}}(C) = \frac{5}{7}$,

$\gamma_{C-\{a_5\}}(C) = \frac{6}{7}, \gamma_{C-\{a_7\}}(C) = \frac{5}{7}$ 。从而得基于优势关系的代数观下

各权重分别为 $\alpha_3 = \frac{2}{7}, \alpha_4 = \frac{2}{7}, \alpha_5 = \frac{1}{7}, \alpha_7 = \frac{2}{7}$ 。

由定义 2 和式(3)、式(4)可得:

$$GK(C) = \frac{10}{49}, GK(C - \{a_3\}) = \frac{18}{49}, GK(C - \{a_4\}) = \frac{18}{49}$$

$$GK(C - \{a_5\}) = \frac{12}{49}, GK(C - \{a_7\}) = \frac{11}{49}, GK(C - \{a_3, a_4\}) = \frac{11}{49}$$

$$GK(C - \{a_7\}) = \frac{12}{49}, Sig(a_3, C - \{a_3\}, C) = \frac{8}{49}$$

$$Sig(a_5, C - \{a_5\}, C) = \frac{1}{49}, Sig(a_7, C - \{a_7\}, C) = \frac{2}{49}$$

从而得到基于优势关系的粒度观下各权重分别为:

$$\beta_3 = \frac{8}{13}, \beta_4 = \frac{2}{13}, \beta_5 = \frac{1}{13}, \beta_7 = \frac{2}{13}$$

根据组合赋权的特征可知:

$$\frac{2}{7} \leq w_3 \leq \frac{8}{13}, \frac{2}{13} \leq w_4 \leq \frac{2}{7}, \frac{1}{13} \leq w_5 \leq \frac{1}{7}, \frac{2}{13} \leq w_7 \leq \frac{2}{7}$$

约简后的数据经过无量纲化处理得到规范决策矩阵:

$$D' = \begin{bmatrix} 0.1542 & 0.1292 & 0.1033 & 0.0760 \\ 0.2951 & 0.1174 & 0.1170 & 0.0873 \\ 0.0846 & 0.2428 & 0.2197 & 0.2551 \\ 0.0898 & 0.2183 & 0.2222 & 0.1974 \\ 0.0715 & 0.0784 & 0.1454 & 0.1230 \\ 0.1024 & 0.0928 & 0.0762 & 0.1186 \\ 0.2024 & 0.1211 & 0.1163 & 0.1427 \end{bmatrix}$$

从而得出各方案的理想值和负理想值分别为:

$$D = \{0.2951, 0.2428, 0.2222, 0.2551\}$$

$$d = \{0.0715, 0.0784, 0.0762, 0.0760\}$$

根据模型(6), 得出各属性权重为:

$$w_3 = 0.2857, w_4 = 0.2857, w_5 = 0.1429, w_7 = 0.2857$$

6 结束语

经典粗糙集理论基于等价关系建立, 但对含有数值型的信息系统进行离散化会导致信息丢失。本文将优势关系引入属性权重的确定方法中, 避免了对连续型属性值必须离散化这一步骤, 保证了原数据的信息量。为确定权重的最优解, 将 2 个概率分布的相对熵定义扩展到任意单维向量的相对熵, 并将相对熵视为任意两单维向量的一种距离。以此原理建立单目标优化模型, 并通过一个算例证明了该方法的有效性。下一步研究工作是优势关系拓展为 α -优势关系, 建立概率粗集赋权模型。

参考文献

- [1] 张文修, 梁 怡, 吴伟志. 信息系统与知识发现[M]. 北京: 科学出版社, 2003.
- [2] Zadeh L A. Fuzzy Logic=Computing with Words[J]. IEEE Trans. on Fuzzy Systems, 1996: 4(2): 103-111.
- [3] 杨霖琳, 秦克云, 裴 峥. 不完备信息系统中的不可区分关系[J]. 计算机工程, 2010, 36(13): 4-6.
- [4] Greco S, Matarazzo B, Slowinski R. Rough Approximation by Dominance Relations[J]. International Journal of Intelligent Systems, 2002, 17(2): 153-171.
- [5] 桂现才. 优势关系下序信息系统的信息量与粗糙熵[J]. 计算机工程与设计, 2008, 29(24): 6340-6343.
- [6] Cover T M, Thomas J A. 信息论基础[M]. 阮吉寿, 张 华, 译. 北京: 机械工业出版社, 2005.
- [7] 陶志富. 几类不确定性决策问题中的熵值理论及其应用[D]. 合肥: 安徽大学, 2010.

编辑 金胡考