文章编号: 1001-0920(2007)12-1437-04

基于模糊粗糙集的因素权重分配方法

柳炳祥, 李海林
(景德镇陶瓷学院 信息工程学院, 江西 景德镇 333403)

摘　要: 基于模糊集与粗糙集的融合,提出一种对多因素分配权重的简单方法. 该方法与传统的权重分配方法有明显的区别. 首先利用模糊聚类分析对数据进行聚类,并提取最佳聚类;然后应用粗糙集中属性的重要程度对各个特征因素进行客观地分配权重. 实例分析验证了所提出方法的有效性.

关键词: 模糊集; 粗糙集; 权重分配
中图分类号: TP183　　**文献标识码**: A

Method of factor weights allocation based on combination of fuzzy and rough set

LIU Bing-xiang, LI Hai-lin
(School of Information Engineering, Jingdezhen Ceramic Institute, Jingdezhen 333403, China. Correspondent: LIU Bing-xiang, E-mail: lbx1966@163.com)

Abstract: Based on fuzzy and rough set, a simple method is proposed to allocate the factor weights, which distinguishes from the traditional ones. Fuzzy cluster analysis is used to aggregate data, and the best clustering is extracted. The conception of the weightiness of rough set is applied to allocate the factor weights with objective way. The examples show the effectiveness of the proposed method.

Key words: Fuzzy set; Rough set; Weights allocation

1 引　　言

日常生活中普遍存在受诸多因素影响的事物,若要对该事物作出评价决策,人们不得不考虑各种因素的权重分配. 在决策过程中权重分配是至关重要的,它反映了各因素在决策过程中所占有的地位或作用,直接影响到决策的结果,通常情况下是由具有丰富经验的专家给出权重. 不可否认,这种方法在一定程度上能反映实际情况,而且评价结果与实际相近,但凭经验得出的权重具有一定的主观性,有时不能客观反映实际情况,从而导致评价结果失真. 有一些方法改进了这种主观性很强的传统方法,如加权统计法、频数统计法等,但它们都建立在多组具有主观性的数据基础上统计权重分配,本质上仍存在主观性. 一些文献[1]给出了基于客观性的权重分配方法,具有实用性,但计算过程复杂,而且某些环节存在一定的不足,如选择置信水平 λ 范围的分类时,没有考虑最佳阈值 λ 的确定而产生的最佳分类.

本文提出一种基于模糊集与粗糙集的客观分配权重的方法,克服了当前一些客观分配权重存在的不足,而且计算过程相对简单,权重分配结果也符合实际.

2 模糊集与粗糙集相关理论

2.1 模糊集相关理论[1-5]

传统聚类分析方法是将每个待处理的对象严格地划分到隶属度为 0 或 1 的某个类中,这往往不能满足现实的需要. 因为客观事物的类别并不是十分明确,即存在模糊性,因此用模糊聚类分析的方法来处理不分明的对象是行之有效的,使分类效果更合乎自然,更符合客观实际,这就是模糊聚类.

设 $X = \{x_1, x_2, \cdots, x_n\}$ 为被分类对象的全体,每个对象 x_i 由一组数据 $\{x_{i1}, x_{i2}, \cdots, x_{im}\}$ 共 m 个因素(特征)来表征,从而得到数据矩阵.

聚类分析的一般步骤如下:

Step 1: 数据标准化. 不同数据一般会有不同的量纲,为了使有不同量纲的量也能进行比较,需要对数据作适当的变换. 常用的方法有标准差标准化和

收稿日期: 2006-09-02; **修回日期**: 2007-02-01.
基金项目: 江西省教育厅科研项目(04JJ23).
作者简介: 柳炳祥(1966—),男,江西九江人,教授,博士,从事数据挖掘、决策支持系统的研究; 李海林(1982—),男,福建龙岩人,硕士生,从事数据挖掘、智能控制的研究.

极差标准化,分别为

$$x_{ij}' = \frac{x_{ij} - \mu_j}{s_j}; \tag{1}$$

$$x_{ij}' = \frac{x_{ij} - \mu_j}{S_j}. \tag{2}$$

其中:μ_j 和 s_j 分别为 x_{ij} 的均值和标准差,$S_j = \max\limits_{1\leq i\leq n}\{x_{ij}\} - \min\limits_{1\leq i\leq n}\{x_{ij}\}$, $i = 1,2,\cdots,n$, $j = 1,2,\cdots,m$.

Step2: 建立模糊相似矩阵 R. 依照传统聚类方法确定相似系数,建立模糊相似矩阵,x_i 与 x_j 的相似度 $r_{ij} = R(x_i, x_j)$. 确定 r_{ij} 的方法有多种,可根据问题性质,选取不同的公式[5].

Step3: 建立模糊等价矩阵 R^*. 利用模糊等价闭包法求出等价矩阵.

Step4: 分类. 根据不同的置信水平 λ 得到不同的分类. 由于等价矩阵 R^* 是一个对称方阵,它的行和列的数目为对象的个数 n,每行或每列都代表一个对象与其他对象的等价关系. 根据给定的置信水平 λ,在等价矩阵 R^* 中查找所有 r_{ij}^* 且须满足 $\lambda \leq r_{ij}^* \leq 1$. 将所有满足该不等式的 r_{ij}^* 所在行代表的对象划分为同一类,其余对象划分为各异的类别,最终可得到动态聚类图;也可从模糊相似矩阵 R 出发,利用 Boole 矩阵法[5] 进行基于置信水平 λ 的分类. 因此,置信水平 λ 不同,所产生的分类结果也不同.

在模糊聚类分析中,对各个不同的 λ 可得到不同的分类. 但许多实际问题需要选择某个阈值 λ,确定样本的一个具体分类,这就提出了如何确定最佳阈值的问题. 为了客观地得到阈值,采用 F- 统计量[5] 确定 λ 最佳值比较符合实际. 在聚类分析过程中,应获得原始数据矩阵,利用该矩阵数据求出

$$\bar{x} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m),$$

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik},\ k = 1,2,\cdots,m, \tag{3}$$

其中 $\bar{x}$ 称为总体样本的中心向量.

设对应于 λ 值的分类数为 r,第 j 类的聚类样本数为 n_j,第 j 类的样本记为 $x_1^{(j)}, x_2^{j}, \cdots, x_{n_j}^{(j)}$,第 j 类的聚类中心为 $\bar{x}^{(j)} = (\bar{x}_1^{(j)}, \bar{x}_2^{(j)}, \cdots, \bar{x}_m^{(j)})$,作 F- 统计量,即

$$F = \frac{\sum_{j}^{r} n_j \| \bar{x}^{(j)} - \bar{x} \|^2/(r-1)}{\sum_{j}^{r}\sum_{i}^{n_j} \| x_i^{(j)} - \bar{x}^{(j)} \|^2/(n-r)}, \tag{4}$$

其中 $\| \bar{x}^{(j)} - \bar{x} \| = \sqrt{\sum(\bar{x}_k^{(j)} - \bar{x}_k)^2}$ 为 $\bar{x}^{(j)}$ 与 $\bar{x}$ 的距离. F- 统计量的分子表征类与类之间的距离,分母表征类内样本间的距离. 因此 F 值越大,说明类与类之间的距离越大,表明类与类之间的差异大,分类效果好.

2.2 粗糙集相关理论[1-4]

粗糙集在分类规则中的基本思想是用户指定数据集中某个或多个属性作为分类的决策属性,根据这些属性的不同取值,将数据分成不同的类别,发现并获取分类规则.

定义 1 信息系统

$$S = \langle U, D, V, f \rangle. \tag{5}$$

其中:U 是一个非空对象集合;C 和 D 是对象的属性集合,即条件属性 C 与决策属性 D;V 是属性值的集合;f 是一个信息函数,即 $f: U \times (C \cup D) \rightarrow V$,它指定了 U 中每组对象的属性值.

定义 2 给定知识库 $k = (U, R)$,对于每个子集 $X \subset U$,等价关系 $R \subset \mathrm{ind}(k)$,称 $R_-(X) = \cup\{Y_i \subset | \mathrm{ind}(R): Y_i \subset X\}$ 为 X 的 R 下近似集.

定义 3 两个属性集 C 与 D 之间的依赖程度 $\gamma(C, D)$ 定义为

$$\gamma(C, D) = | \mathrm{POS}_C(D) | / | U |. \tag{6}$$

其中:$\mathrm{POS}_C(D) = C_-(D)$, $| U |$ 表示整个集合对象的个数.

定义 4 属性 $a \in C$,属性 a 关于 D 的重要程度定义为

$$\mathrm{SGF}(a, C, D) = \gamma(C, D) - \gamma(C - \{a\}, D), \tag{7}$$

其中 $\gamma(C - \{a\}, D)$ 表示在 C 中缺少属性 a 后,条件属性对决策属性的依赖程度.

3 基于模糊粗糙集的因素权重分配方法

一般情况下,事物可由诸多因素表征,评价该事物通常出于一个或多个目的(决策属性),评价的因素来自影响该事物的条件属性. 如果已知决策属性和条件属性,则利用粗糙集方法很容易求得各个因素的重要程度,从而解决权重分配问题. 然而,有时对事物的评价,只根据影响它的各因素作出综合评价,决策属性未知,若要确定各因素的权重,就必须从各属性之间的相互关系出发来解决权重分配问题. 利用粗糙集和模糊集的相关理论,首先根据所用的特征属性进行模糊聚类,找到最佳分类,将它看作某种决策属性的分类,可得到按某种假设决策属性的 k 等价集. 根据同样的方法,依次删除单个属性 C_i 再进行模糊聚类,从而得到条件属性 $C - \{c_i\}$ 的等价集;然后计算条件属性与决策属性之间的依赖程度 $\gamma(C - \{c_i\}, D)$,求得属性 c_i 的重要程度;最后利用归一化重要程度的方法来求解各因素的权重. 具体步骤如下:

Step1: 对象集为 $X = \{x_1, x_2, \cdots, x_n\}$,条件属性值为 $(x_{i1}, x_{i2}, \cdots, x_{im})$,得到原始数据矩阵

$$A = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \ddots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix},$$

按模糊聚类分析的一般步骤进行分类.

Step2: 由 F- 统计量方法确定最佳置信水平阈值 λ,找出最佳分类

$$Y = \{Y_1, Y_2, \cdots, Y_s\}, \tag{8}$$

Y_i 表示一个等价集. 将该分类当作某种决策属性的等价集的集合.

Step3: 删除条件属性 c_i $(i = 1, 2, \cdots, m)$ 后,得到删除后的原始数据矩阵. 对该矩阵按模糊聚类分析的方法进行分类,利用 F- 统计量确定最佳置信水平,找出删除条件属性 c_i 后的最佳分类,得到依次删除 c_i 的分类集

$$E = \{E_1, E_2, \cdots, E_m\}. \tag{9}$$

其中:对于不同的 i 值,k 也可以不同; $E_i = \{Y_1^{(i)}, Y_2^{(i)}, \cdots, Y_k^{(i)}\}$ 表示删除第 i 个条件属性后得到的分类等价集; $Y_l^{(i)}$ $(1 \leq l \leq k)$ 表示删除第 i 个条件属性后所得到分类的第 l 个等价集.

Step4: 利用粗糙集相关原理,求解每个属性的重要程度. 分别求解决策属性的各等价集的下近似集的并集,公式为

$$\mathrm{POS}_{C-\{c_i\}}(D) = \{C - \{c_i\}\}_-(D) = \cup\{\{C - \{c_i\}\}_- Y_l\}, \tag{10}$$

其中 $1 \leq l \leq s$,且由条件属性 $C - c_i$ 决定分类的等价集为 $E_i = \{Y_1^{(i)}, Y_2^{(i)}, \cdots, Y_k^{(i)}\}$. 由粗糙集定义 3 计算两个属性集的依赖程度

$$\gamma(C - \{c_i\}, D) = |\mathrm{POS}_{C-\{c_i\}}(D)| / |U|. \tag{11}$$

再由粗糙集相关理论定义 4 求解条件属性 c_i 的重要程度 $\mathrm{SGF}(c_i, C, D)$.

Step5: 根据每个条件属性的重要程度,用归一化处理方法分配权重,权重分配公式为

$$W_i = \mathrm{SGF}(c_i, C, D) / \sum_{k=1}^{m} \mathrm{SGF}(c_k, C, D). \tag{12}$$

通过数据之间的信息关系,利用上述方法客观地分配权重,这就是本文提出的一种客观分配权重的方法.

4　实　　例

用上述方法对文献[1]的知识表达系统的 5 个因素进行权重分析,具体知识表达系统如表 1 所示.

表 1　某行业指标数据

编号	条件属性				
	因素 A	因素 B	因素 C	因素 D	因素 E
1	0.88	130	410	1.43	2.98
2	4.33	21	180	3.38	2.40
3	4.91	50	969	5.42	1.46
4	16.20	26	1 020	5.16	1.16
5	15.38	87	1 540	4.40	0.65
6	14.56	140	2 270	4.34	0.27
7	77.70	135	2 140	6.69	0.36
8	82.10	332	2 660	14.60	0.49
9	95.94	136	2 230	10.18	0.37
10	202.10	408	6 790	8.86	0.31
11	262.40	500	16 050	13.60	0.15
12	185.10	670	7 200	14.80	0.26

根据上述权重分配方法和 F- 统计量方法,通过计算得到 12 个样本的最佳分类,当置信水平为 $\lambda = 0.70$ 时,分类情况为 $\{1\}, \{2\}, \{3, 4\}, \{5, 6, 7, 9\}, \{8\}, \{10, 11, 12\}$,并将这种分类当作某种决策属性 D 的等价集,故认为可以将知识表达系统按决策属性分成 6 类. 在文献[1]中,通过主观方式给出了置信水平 λ 的取值范围 (λ 的取值分界点为 0.65, 0.70, 0.75, 0.80, 0.85),利用这些区间实现对象的硬划分,导致客观分配权重的方法趋于主观化,最终影响权重值的客观分配. 本文利用 F- 统计量方法确定最佳阈值 λ,根据 λ 值划分对象所得到的等价类作为决策属性的等价集,这样就客观地考虑了各因素关系,并将全部因素之间的关系作为决策属性,进而充分地挖掘出各对象之间的数据关系,为进行客观权重分配作准备.

按照 Step3,找出删除 c_i 条件属性的分类,同样用模糊聚类的方法进行分类. 经计算,各种最佳的分类情况如下:

删除因素 A 后的分类为 $\{1, 2, 3, 4\}, \{5, 6, 7, 8\}, \{9\}, \{10\}, \{11, 12\}$;

删除因素 B 后的分类为 $\{1\}, \{2, 3, 4\}, \{5, 6\}, \{7\}, \{8, 9\}, \{10\}, \{11, 12\}$;

删除因素 C 后的分类为 $\{1\}, \{2, 3, 4\}, \{5, 6\}, \{7\}, \{8\}, \{9\}, \{10\}, \{11, 12\}$;

删除因素 D 后的分类为 $\{1\}, \{2, 3, 4\}, \{5, 6\}, \{7, 8, 9\}, \{10, 11, 12\}$;

删除因素 E 后的分类为 $\{1\}, \{2, 3, 4, 5, 6\}, \{7\}, \{8\}, \{9\}, \{10\}, \{11, 12\}$.

根据 Step4,因素 A 的重要程度为

$$\mathrm{SGF}(A, C', D') = 1 - \gamma(C' - c_i, D') = 1 - 4/12 = 2/3.$$

同理,其他各因素的重要程度分别为

$$\mathrm{SGF}(B, C', D') = 5/12;$$

$$SGF(C, C', D) = 1/4;$$
$$SGF(D, C', D) = 1/2;$$
$$SGF(E, C', D) = 5/12.$$

最后由归一化处理方法得到各个因素的权重,分别为

$$(w_A, w_B, w_C, w_D, w_E) = (0.294\,38, 0.185\,549, 0.111\,508, 0.223\,015, 0.185\,549).$$

客观地进行因素权重分配的关键在于如何处理数据之间的相互关系,并根据这些数据关系值进一步确定各因素的重要程度,进而客观地得到各因素的权重.由上述数据分析过程可知,算法利用了模糊聚类分析方法和粗糙集的相关原理,由于模糊数学所研究的对象与对象之间的关系具有模糊性,粗糙集是处理模糊性与不确定性的有利工具,而且在研究客观分配因素权重过程中,最主要是要挖掘出数据集合之间的信息关系,此过程具有很强的模糊性和不确定性.因此,该方法通过利用模糊数学和粗糙集的相关原理来解决客观分配因素权重的不确定性问题是行之有效的方法,其结果与文献[1]相比,更具有客观性和实用性.

5 结 语

从实例分析可知,基于模糊粗糙集的因素权重分配方法完全依靠数据之间的信息关系来确定权重,它不存在专家评估的主观性,为知识表达系统进一步作出决策打下基础,具有实用性.文献[1]因为没有考虑分类的最佳选择,从而影响了权重的计算,导致最后算出的权重与该文献的结果不一致,但总体权重分配还是相似的,即因素 A 与因素 D 的权重总体上大于其他几个因素的权重.从实例中的权重数值可以看出,因素 B 与因素 E 的权重相等,从数据信息关系分析出发,说明这 2 个因素的权重必须一样,如果让具有丰富经验的专家给出权重值,同样会给出权重相等的可能,因此是符合实际的.

参考文献(References)

[1] 黄定轩. 基于客观信息熵的多因素权重分配法[J]. 系统工程理论方法应用, 2002, 12(4): 321-324.
(Huang Ding-xuan. Means of weights allocation with multi-factors based on impersonal message entropy[J]. Systems Engineering — Theory Methodology Applications, 2002, 12(4): 321-324.)

[2] 陈文伟. 数据仓库与数据挖掘[M]. 北京: 人民邮电出版社, 2004: 137-143.
(Chen Wen-wei. Data warehouse and data mining[M]. Beijing: People's Posts and Telecommunications Publishing House, 2004: 137-143.)

[3] 梁循. 数据挖掘算法与应用[M]. 北京: 北京大学出版社, 2006: 219-236.
(Liang Xun. Data-mining algorithms and applications [M]. Beijing: Peking University Press, 2006: 219-236.)

[4] 曾黄麟. 智能计算[M]. 重庆: 重庆大学出版社, 2004: 14-156.
(Zeng Huang-lin. Intelligence computing [M]. Chongqing: Chongqing University Press, 2004: 14-156.)

[5] 谢季坚. 模糊数学方法及其应用[M]. 武汉: 华中科技大学出版社, 2000: 81-118.
(Xie Ji-jian. Fuzzy mathematics method and applications [M]. Wuhan: Press of Huazhong University of Science and Technology, 2000: 81-118.)

(上接第 1436 页)

[7] 周锐. 基于模糊逻辑的导弹复合控制系统优化设计[J]. 控制与决策, 2006, 21(7): 825-828.
(Zhou Rui. Design of missile blended control system based on fuzzy logic[J]. Control and Decision, 2006, 21(7): 825-828.)

[8] Christopher D C. Failure recovery in redundant serial manipulators[D]. Austin: The University of Texas at Austin, 2000.

[9] 姜力, 蔡鹤皋, 刘宏. 基于滑模位置控制的机器人灵巧手模糊自适应阻抗控制[J]. 控制与决策, 2001, 16(5): 612-616.
(Jiang Li, Cai He-gao, Liu Hong. Fuzzy adaptive impedance control of dextrous robot hand based on sliding mode position control[J]. Control and Decision, 2001, 16(5): 612-616.)

[10] Wang H L, Li J W, Liu H. Practical limitations of an algorithm for the singular value decomposition as applied to redundant manipulators[C]. IEEE Conf on Robotics, Automation and Mechatronics. Bangkok, 2006:1-6.)