

文章编号: 1001-0920(2011)01-0153-04

基于信息熵的专家聚类赋权方法

周璇, 张凤鸣, 惠晓滨, 李克武

(空军工程大学 工程学院, 西安 710038)

摘要: 鉴于群组决策专家赋权方法研究中, 现有赋权方法虽然考虑了专家给出的排序向量的一致性, 但缺乏对排序向量信息相似性的度量, 导致可能出现排序向量与群体共识相近, 但信息不确定性较大的专家被赋予了与其他专家相同权重的问题. 基于此, 提出一种基于信息熵的专家聚类赋权方法, 运用信息相似系数对排序向量进行聚类分析, 根据聚类结果和排序向量的信息熵来确定专家的权重. 具体算例表明, 该方法有效且可行.

关键词: 熵; 信息相似系数; 聚类分析

中图分类号: O236

文献标识码: A

Method for determining experts' weights based on entropy and cluster analysis

ZHOU Xuan, ZHANG Feng-ming, HUI Xiao-bin, LI Ke-wu

(College of Engineering, Air Force Engineering University, Xi'an 710038, China. Correspondent: ZHOU Xuan, E-mail: zhouxuan_333@126.com)

Abstract: According to the methods of determining experts' weights in group decision-making, the existing methods take into account the consistency of experts' collating vectors, but it is lack of the measure of its information similarity. So it may occur that although the collating vector is similar to the group consensus, information uncertainty is great of a certain expert. However, it is given the same weight to the other experts. For this, a method for deriving experts' weights based on entropy and cluster analysis is proposed, in which the collating vectors of all experts are classified with information similarity coefficient, and the experts' weights are determined according to the result of classification and entropy of collating vectors. Finally, a numerical example shows that the method is effective and feasible.

Key words: entropy; information similarity coefficient; cluster analysis

1 引言

在多属性、多目标决策过程中, 为了克服单个专家主观偏见和模糊认识的影响, 通常采用专家群组决策的方法来提高决策的正确性. 目前, 在群组决策过程中基于判断矩阵的专家赋权方法大致可分为两类: 一类是根据专家给出的判断矩阵的一致性程度来赋权, 判断矩阵的一致性程度越好, 相应的专家权重越大; 另一类是利用系统聚类分析的思想, 对专家先进行分类, 再赋权^[1]. 传统的聚类方法有基于距离的、密度的、网格的和相似系数的等. 文献[2]提出了一种基于相似系数的专家聚类赋权方法, 以排序向量相容度为测度, 根据专家排序向量的相似程度为专家聚类, 然后根据专家判断矩阵的一致性程度为专家赋权. 传统的基于向量相似系数的聚类算法在描述物理位置接近的点时起到了重要作用, 但它们对整体相似程度

较高, 而个别元素波动较大的向量之间的相似性度量却发挥不了应有的效果. 因为它们没有考虑聚类对象实际含有信息量的大小和信息的相似程度, 导致在专家聚类赋权方法中可能出现虽然评判向量与群体共识相近, 但信息不确定性大的专家被赋予了与其他专家相同权重的问题. 根据信息分类^[3]的要求, 被聚成类的一组对象不仅应具有极高的相似性, 而且应反映同一个信息, 信息的相似程度要高.

基于以上考虑, 本文提出一种基于信息熵的专家聚类赋权方法, 利用信息相似系数度量专家意见的相似程度, 不仅运用熵^[3-10]的原理对专家进行聚类分析, 而且运用信息熵为类中专家赋权. 使得在整体评价中, 排序向量信息相似系数小、共识较好、专家数目的类别间具有更高的权重. 在同一信息相似度内的专家群体中, 逻辑清晰、思维严密、信息熵少的专家

收稿日期: 2009-11-02; 修回日期: 2009-12-28.

作者简介: 周璇(1986—), 男, 博士生, 从事信息系统工程与智能决策等研究; 张凤鸣(1963—), 男, 教授, 博士生导师, 从事智能信息处理、智能信息系统和智能决策等研究.

拥有更高的类内权重,从而实现少数服从多数,精确胜过模糊的原则.

2 基于信息熵的专家聚类赋权方法

基于信息熵的专家聚类赋权方法分两步:首先根据各专家排序向量的信息相似系数对专家进行聚类;然后根据聚类结果中的专家人数和排序向量的信息熵为各专家赋权.

2.1 专家聚类分析

对专家进行聚类必须根据专家给出的排序向量元素的特点定义一种合理的相似性测度.为了衡量各排序向量所含信息的相似性,采用信息相似系数^[3]作为聚类的标准.

假设有 m 个专家在某一准则下对 n 个评价对象进行评价,第 i 个专家给出的判断矩阵为 A_i ,由判断矩阵 A_i 求出的评价对象排序向量为 $U_i = (u_{i1}, u_{i2}, \dots, u_{in})^T$, $i = 1, 2, \dots, m$. 由此,可以得到每位专家给出的评价对象的状态空间为

$$\begin{bmatrix} X \\ U_i \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ u_{i1} & u_{i2} & \cdots & u_{in} \end{bmatrix}, \quad (1)$$

其中符号集合 $X = (x_1, x_2, \dots, x_n)$ 表示评价对象的集合.

为了对专家给出的信息进行分类,定义离散量为

$$D(U_i) \triangleq D(u_{i1}, u_{i2}, \dots, u_{in}) = \left(\sum_{j=1}^n u_{ij} \right) \log_2 \left(\sum_{j=1}^n u_{ij} \right) - \sum_{j=1}^n u_{ij} \log_2(u_{ij}). \quad (2)$$

由于 $\sum_{j=1}^n u_{ij} = 1$, 可得

$$D(U_i) = - \sum_{j=1}^n u_{ij} \log_2(u_{ij}).$$

为了衡量专家给出排序向量的信息相似程度,采用信息相似性系数 $\Delta(U_i, U_j)$ 进行度量,其中 U_i 和 U_j 分别为专家 i 和 j 给出的评价对象的状态空间,用符号可表示为 $U_i : [u_{i1}, u_{i2}, \dots, u_{in}]$ 和 $U_j : [u_{j1}, u_{j2}, \dots, u_{jn}]$, 则其和的评价对象的状态空间为

$$U_i + U_j : [u_{i1} + u_{j1}, u_{i2} + u_{j2}, \dots, u_{in} + u_{jn}]. \quad (3)$$

其对应的离散量分别为 $D(U_i)$, $D(U_j)$, $D(U_i + U_j)$. 经过证明可以发现, $D(U_i + U_j) \geq D(U_i) + D(U_j)$, 当 U_i 和 U_j 线性相关且完全相似时取等号, 所以定义离散量之间的差值为信息相似性系数 $\Delta(U_i, U_j)$, 有

$$\Delta(U_i, U_j) = D(U_i + U_j) - D(U_i) - D(U_j) = \left[\sum_{k=1}^n (u_{ik} + u_{jk}) \right] \log_2 \left[\sum_{k=1}^n (u_{ik} + u_{jk}) \right] -$$

$$\begin{aligned} & \sum_{k=1}^n (u_{ik} + u_{jk}) \log_2(u_{ik} + u_{jk}) - \\ & \left(\sum_{k=1}^n u_{ik} \right) \log_2 \left(\sum_{k=1}^n u_{ik} \right) + \sum_{k=1}^n u_{jk} \log_2(u_{jk}) - \\ & \left(\sum_{k=1}^n u_{jk} \right) \log_2 \left(\sum_{k=1}^n u_{jk} \right) + \sum_{k=1}^n u_{jk} \log_2(u_{jk}) = \\ & 2 - \sum_{k=1}^n [(u_{ik} + u_{jk}) \log_2(u_{ik} + u_{jk}) - \\ & u_{ik} \log_2(u_{ik}) - u_{jk} \log_2(u_{jk})] = \\ & 2 - \sum_{k=1}^n D(u_{ik} + u_{jk}). \end{aligned} \quad (4)$$

由式(4)可得

$$\Delta(U_i, U_j) = \Delta(U_j, U_i), \Delta(U_i, U_j) \in [0, 2].$$

当 U_i 和 U_j 完全相似时, $\Delta(U_i, U_j) = 0$. 当 $\Delta(U_i, U_j)$ 达到最大值 2 时, 表示 U_i 和 U_j 完全不相似. 为了使信息相似系数在 $[0, 1]$ 内取值, 定义标准信息相似系数为

$$I(U_i, U_j) = \Delta(U_i, U_j)/2. \quad (5)$$

利用 $I(U_i, U_j)$ 的大小来衡量排序向量之间的信息相似程度, 确定聚类规则. 聚类分析步骤如下:

Step 1: 根据专家给出的排序向量, 按照式(1)构造每位专家给出的评价对象的状态空间.

Step 2: 根据式(3)构造两两专家和的评价对象的状态空间, 并按照式(4)和(5)计算两两专家排序向量之间的信息相似系数和标准信息相似系数.

Step 3: 以标准信息相似系数最小(即 $\min I(U_i, U_j)$) 和 $I(U_i, U_j) \leq k$ 为准则, 确定两两专家聚类的集合, 其中 k 为依据实际聚类需要而确定的信息相似阈值.

Step 4: 对两两专家聚类集合中的含有相同专家的子集合进行“并”操作, 确定最终的专家聚类结果.

2.2 专家赋权分析

为专家赋权, 不仅要考虑专家所在类中专家的人数, 还要考虑专家给出的排序向量所蕴含的信息量. 对于类容量大的类, 其专家给出的评价信息代表了大多数专家的意见, 对应专家应赋予较大的权值^[2]. 对于类中专家权值, 传统的赋权方法有平均赋权法和基于判断矩阵一致性的赋权方法, 但是均存在缺陷: 平均法没有考虑排序向量之间的差异性; 基于判断矩阵一致性的赋权法虽然考虑了判断矩阵的一致性, 但没有考虑排序向量所表达的信息量和不确定程度, 导致可能出现一致程度高, 而排序向量信息不确定性大的专家却被赋予高权值的问题. 熵作为信息不确定性的度量, 如果专家给出的排序向量的信息熵越小, 说明该专家的思路越清晰、不确定因素越小, 相应地该专

家应赋予更大的权值. 因此, 为了尽量减少专家信息的不确定性, 类间权值用每类中专家人数进行测度, 类内权值用信息熵进行测度. 对于只有一位专家的类别, 规定该专家的类内权值为1.

假设 m 位专家被分为 l 类, 第 i 类中专家的数目为 Φ_i , λ_i 为类间权重, α_{ij} 为第 i 类中第 j 位专家的类内权重, ω_{ij} 为第 i 类中第 j 位专家的总体权重, $U_{ij} = (u_{ij1}, u_{ij2}, \dots, u_{ijn})^T$ 为第 i 类中第 j 位专家给出的排序向量, U_{ij} 的信息熵为 $H(U_{ij})$, 其中 $i = 1, 2, \dots, l, j = 1, 2, \dots, \Phi_i$.

归纳上述推理可以得到 $H(U_{ij})$, λ_i 和 α_{ij} 分别为

$$H(U_{ij}) = -\frac{1}{\log_2 n} \sum_{k=1}^n u_{ijk} \log_2(u_{ijk}) = \frac{D(U_{ij})}{\log_2 n}, \quad (6)$$

$$\lambda_i = \Phi_i^2 / \sum_{j=1}^l \Phi_j^2, \quad (7)$$

$$\alpha_{ij} = (1 - H(U_{ij})) / \sum_{k=1}^{\Phi_i} [1 - H(U_{ik})]. \quad (8)$$

通过求得的 λ_i 和 α_{ij} , 可以对专家的总体权值 ω_{ij} 进行集结, 即

$$\omega_{ij} = \alpha_{ij} \lambda_i. \quad (9)$$

3 算例分析

本文引用文献[2]的算例, 假设有7位专家在某一准则下对3个对象进行评价, 他们所列出的判断矩阵分别为

$$E_1 = \begin{bmatrix} 1 & 4 & 7 \\ 1/4 & 1 & 3 \\ 1/7 & 1/3 & 1 \end{bmatrix}, E_2 = \begin{bmatrix} 1 & 4 & 5 \\ 1/4 & 1 & 4 \\ 1/5 & 1/4 & 1 \end{bmatrix},$$

$$E_3 = \begin{bmatrix} 1 & 2 & 5 \\ 1/2 & 1 & 3 \\ 1/5 & 1/3 & 1 \end{bmatrix}, E_4 = \begin{bmatrix} 1 & 5 & 8 \\ 1/5 & 1 & 4 \\ 1/8 & 1/4 & 1 \end{bmatrix},$$

$$E_5 = \begin{bmatrix} 1 & 2 & 6 \\ 1/2 & 1 & 5 \\ 1/6 & 1/5 & 1 \end{bmatrix}, E_6 = \begin{bmatrix} 1 & 5 & 6 \\ 1/5 & 1 & 3 \\ 1/6 & 1/3 & 1 \end{bmatrix},$$

$$E_7 = \begin{bmatrix} 1 & 2 & 5 \\ 1/2 & 1 & 2 \\ 1/5 & 1/2 & 1 \end{bmatrix}.$$

运用文献[2]提供的方法, 可以求得各专家给出的评价对象排序向量为

$$\begin{aligned} U_1 &= (0.71, 0.21, 0.08)^T, U_2 = (0.66, 0.25, 0.09)^T, \\ U_3 &= (0.58, 0.31, 0.11)^T, U_4 = (0.73, 0.20, 0.07)^T, \\ U_5 &= (0.58, 0.34, 0.08)^T, U_6 = (0.72, 0.19, 0.09)^T, \\ U_7 &= (0.60, 0.28, 0.12)^T. \end{aligned}$$

首先, 根据专家给出的排序向量进行聚类分析.

由于专家评价的对象很少, 信息相似程度要求较高, 取信息相似阈值 k 为 0.001. 通过计算可以得到专家两两聚类的集合为 $\{(1, 4), (3, 7), (4, 1), (6, 1), (7, 3)\}$, 对含有相同专家的子集合进行并操作, 可以得到最终的聚类结果为 $\{(1, 4, 6), (3, 7), (2), (5)\}$. 与文献[2]聚类结果基本相同, 专家被分为4类, 唯一的区别是专家5被独自分为一类. 虽然 U_5 , U_3 和 U_7 整体相似, 利用相容度测度满足聚类的要求, 但 U_5 较 U_3 和 U_7 中的元素波动较大, 造成了信息相似性差, 没有与 U_3 和 U_7 表达同样的评价信息, 不符合信息分类的要求. 因此, 专家5应自成一类.

其次, 根据每类中专家人数和信息熵为专家进行赋权分析. 根据式(6)可以得到每位专家排序向量的信息熵为

$$H(U_1) = 0.70, H(U_2) = 0.76, H(U_3) = 0.84,$$

$$H(U_4) = 0.67, H(U_5) = 0.81, H(U_6) = 0.70,$$

$$H(U_7) = 0.82.$$

根据式(7)可以得到类间权重为

$$\lambda_1 = \frac{3^2}{3^2 + 2^2 + 1 + 1} = 9/15,$$

$$\lambda_2 = 4/15, \lambda_3 = \lambda_4 = 1/15.$$

根据式(8)可以得到类内权重为

$$\alpha_{11} = 0.32, \alpha_{12} = 0.36, \alpha_{13} = 0.32,$$

$$\alpha_{21} = 0.49, \alpha_{22} = 0.51, \alpha_{31} = \alpha_{41} = 1.$$

最后, 利用式(9)对专家权值进行集结, 可以得到专家总体权值向量为

$$\varpi = (0.19, 0.07, 0.13, 0.22, 0.07, 0.19, 0.13)^T.$$

综合7位专家的权值和个人排序向量, 可得最终的综合排序向量为

$$U = (0.67, 0.24, 0.09)^T.$$

从上述分析过程可以看出, 在专家人数多的类别中, 信息熵为专家赋权起到了决定性的作用, 信息熵越小被赋予的权值越大, 从而对整体排序向量的影响程度越大. 对于人数较少的类别, 即使信息熵很小也被赋予了较小的权值, 有效避免了个人主观意见的影响, 体现了少数服从多数、精确胜过模糊的原则, 符合决策的实际需要.

4 结 论

本文根据对现有专家聚类赋权方法的分析, 结合信息熵理论, 提出了一种基于信息熵的聚类赋权方法, 根据信息熵为专家聚类和赋权, 从而为专家聚类赋权提供了新的思路, 能够减少决策过程中的不确定性. 算例分析表明, 该方法可行有效, 有利于提高决策的正确性.

参考文献(References)

- [1] 曾雪兰, 吉建华, 吴小欢. 基于相容性指标的聚类分析专家赋权法[J]. 广西大学学报, 2005, 30(4): 337-340.
(Zeng X L, Ji J H, Wu X H. Research on method for experts' weights in group decision making based on reciprocal judgment matrix[J]. J of Guangxi University, 2005, 30(4): 337-340.)
- [2] 高阳, 罗贤新, 胡颖. 基于判断矩阵的专家聚类赋权研究[J]. 系统工程与电子技术, 2009, 31(3): 593-596.
(Gao Y, Luo X X, Hu Y. Research on methods for deriving experts' weights based on judgment matrix and cluster analysis[J]. Systems Engineering and Electronics, 2009, 31(3): 593-596.)
- [3] 傅祖芸. 信息论[M]. 北京: 电子工业出版社, 2007.
(Fu Z Y. Information theory[M]. Beijing: Publishing House of Electronics Industry, 2007.)
- [4] Ding S F, Shi Z Z. Studies on incidence pattern recognition based on information entropy[J]. J of Information Science, 2005, 31(6): 497-502.
- [5] 杨明, 杨萍. 垂直分布多决策表下基于条件信息熵的近似约简[J]. 控制与决策, 2008, 23(10): 1103-1108.
(Yang M, Yang P. Approximate reduction based on conditional information entropy over vertically partitioned multi-decision table[J]. Control and Decision, 2008, 23(10): 1103-1108.)
- [6] 汪泽焱, 益晓新. 基于基点和熵的系统评价指标赋权法[J]. 解放军理工大学学报, 2002, 3(3): 92-95.
(Wang Z Y, Yi X X. Method of determining the index weight based on base points and entropy[J]. J of PLA University of Science and Technology, 2002, 3(3): 92-95.)
- [7] Cover T M, Thomas J A. Elements of information theory[M]. New York: John Wiley and Sons, 1991.
- [8] Hwang C L, Yoon K S. Multiple attribute decision making and application[M]. New York: Springer-Verlag, 1981.
- [9] Jessop A. Entropy in multi-attribute problems[J]. J of Multi-criteria Decision Analysis, 1999, 8(2): 61-70.
- [10] 赵玉普, 冯书兴, 王佳, 等. 信息熵空间作战指挥多属性决策研究[J]. 火力与指挥控制, 2008, 33(9): 55-59.
(Zhao Y P, Feng S X, Wang J, et al. Research on space operation command multi-attribute decision-making based on information entropy[J]. Fire Control and Command Control, 2008, 33(9): 55-59.)

(上接第152页)

- [8] 唐发明, 王仲东, 陈绵云. 支持向量机多类分类算法研究[J]. 控制与决策, 2005, 20(7): 746-754.
(Tang F M, Wang Z D, Chen J Y. On multiclass classification methods of support vector machines[J]. Control and Decision, 2005, 20(7): 746-754.)
- [9] Sahbi Hichem, Geman Donald, Perona Pietro. A hierarchy of support vector machines for pattern detection[J]. J of Machine Learning Research, 2006, 7(10): 2087-2123.
- [10] 唐发明, 王仲东, 陈绵云. 一种新的二叉树多类支持向量机算法[J]. 计算机工程与应用, 2005, 41(7): 24-26.
(Tang F M, Wang Z D, Chen M Y. An improved multiclass support vector machines based on binary tree[J]. Computer Engineering and Applications, 2005, 41(7): 24-26.)
- [11] 张晓平, 杨洁明. 一种新的支持向量机多类分类二叉树生成算法[J]. 机器工程与自动化, 2007, (3): 1-3.
(Zhang X P, Yang J M. A new algorithm of generating binary tree on multiclass classification methods for support vector machines[J]. Mechanical Engineering and Automation, 2007, (3): 1-3.)
- [12] 赵晖, 荣莉莉, 李晓. 一种设计层次支持向量机多类分类器的新方法[J]. 计算机应用研究, 2006, 23(6): 34-37.
(Zhao H, Rong L L, Li X. New method of design hierarchical support vector machine multi-class classifier[J]. Application Research of Computers, 2006, 23(6): 34-37.)
- [13] 谢志强, 高丽, 杨静. 基于球结构的完全二叉树SVM多类分类算法[J]. 计算机应用研究, 2008, 25(11): 3268-3274.
(Xie Z Q, Gao L, Yang J. SVM multi-class classification algorithm based on full-binary tree of sphere-structured[J]. Application Research of Computers, 2008, 25(11): 3268-3274.)
- [14] 夏思宇, 潘泓, 金立左. 非平衡二叉树多类支持向量机分类方法[J]. 计算机工程与应用, 2009, 45(17): 167-169.
(Xia S Y, Pan H, Jin L Z. Multi-class SVM method based on non-balanced binary tree[J]. Computer Engineering and Applications, 2009, 45(17): 167-169.)
- [15] 朱远平, 戴汝为. 基于SVM决策树的文本分类器[J]. 模式识别与人工智能, 2005, 18(4): 412-416.
(Zhu Y P, Dai R W. Text classifier based on SVM decision tree[J]. Pattern Recognition and Artificial Intelligence, 2005, 18(4): 412-416.)