文章编号: 1001-0920(2007)05-0515-05

密度加权平均中间算子及其在多属性决策中的应用

易平涛, 郭亚军, 张丹宁
(东北大学 工商管理学院, 沈阳 110004)

摘　要: 针对已有的信息集结算子没有考虑属性信息分布疏密程度的情况,研究了一维数据聚类的问题,并给出一种一维聚类的方法. 在此基础上,开发了密度加权平均(DWA)中间算子,并将该算子与已有的5种信息集结算子进行合成,得到的合成算子兼顾了多种算子的特点. 最后通过一个算例验证了DWA算子及合成算子的一些新颖的特征.

关键词: 一维聚类; 密度加权平均中间算子; 多属性决策
中图分类号: C934　　**文献标识码**: A

Density weighted averaging middle operator and application in multi-attribute decision making

YI Ping-tao, GUO Ya-jun, ZHANG Dan-ning
(College of Business Administration, Northeastern University, Shenyang 110004, China. Correspondent: YI Ping-tao, E-mail: yipingtao@163.com)

Abstract: Aiming at the condition that existing information aggregation operators neglect the density degree of attribute information distribution, the theories and a method of one-dimensional clustering are introduced. Based on above theories, the density weighted averaging (DWA) middle operator is developed. Composite operators composed of the DWA and other five existing operators give attention to manifold operators' characteristics. Finally, an example shows some novel characteristics of the DWA and composite operators.

Key words: One-dimensional clustering; Density weighted averaging middle operator; Multi-attribute decision making

1 引　言

目前,有关决策信息集结算子的研究成果比较丰富. 相对于常用的加权算术平均(WAA)、加权几何平均(WGA)等算子,Yager提出的有序加权平均(OWA)算子是一种独特的多属性信息集结方法[1]. 它主要用于解决属性权重信息未知情形下的多源信息融合问题,已在管理决策、专家系统、人工智能、模糊系统等多个领域取得了广泛的应用[2,3]. 后续的理论研究主要从两方面展开:一是OWA算子中属性权重的确定方法[4,5];二是OWA算子与其他算子的融合及OWA算子在不确定性或模糊语言环境下的拓展[6-10].

与WAA(或WGA)算子对各个属性的重要性加权的思路不同,OWA算子是对方案的属性值按照所处的大小位置进行加权,运用了属性值之间大小秩序的信息,并通过位置权重的配置体现与、或、偏与、偏或等集结特征,但是没有考虑属性值之间分布的疏密程度. 在多属性决策中,属性值分布几乎都是不均匀的,此时考虑数据疏密程度的信息显得十分必要. 在一组数据中,数据越集中,说明信息的一致性程度越高;数据越分散,说明信息的一致性程度越低. 决策者可以偏好集中的信息(强调群体意见),也可以偏好分散的信息(强调个别意见),通过对数据集中数据分布特征的判别,实现数据组在信息分布上的与、或、偏与、偏或等集结运算.

本文研究一维数据聚类问题,并提出一种简洁有效的一维聚类方法;在此基础上,开发了基于属性值分布疏密程度信息的密度加权平均(DWA)中间算子. 该算子与算术平均(AA)、加权算术平均(WAA)、有序加权算术平均(OWA)、最小(Min)、最大(Max)等算子结合,能利用更多的决策信息,从而达到更好的决策效果.

收稿日期: 2006-04-10; 修回日期: 2006-06-05.
基金项目: 国家自然科学基金项目(70472032).
作者简介: 易平涛(1981 —),男,湖南永州人,讲师,博士,从事决策理论与技术等研究; 郭亚军(1952 —),男,辽宁开原人,教授,博士生导师,从事系统评价、决策支持等研究.

2 一维数据聚类问题及方法

一维数据聚类是指对一维数轴上的若干数据点按照间隔疏密程度进行分组的问题. 为方便起见,记 $A = \{a_i \mid a_i \in R, i = 1,2,\cdots,n\}$, $M = \{1,2,\cdots,m\}$, $N = \{1,2,\cdots,n\}$.

定义1 对于数据集 A, $A_1, A_2, \cdots, A_m$ 为 A 的非空子集合,若满足

$$A_i \cap A_j = \varnothing, i \neq j, i, j \in M;$$

$$A_1 \cup A_2 \cup \cdots \cup A_m = A.$$

则称 $A_1, A_2, \cdots, A_m$ 为 A 的一个划分.

定义2 对于数据集 A,称 $D(A) = \max(A) - \min(A)$ 为 A 的极差.

定义3 对于 A 的一个划分 $A_1, A_2, \cdots, A_m$,称 $D'(A_1, A_2, \cdots, A_m) = \sum_{i=1}^{m} D(A_i)$ 为类总极差. 显然,对于任意一个集合 $A_i (i \in M)$,有 $D'(A_i) = D(A_i)$.

定义4 $A_1, A_2, \cdots, A_m$ 为数据集 A 的任意一个划分,若满足 $D'(A_1', A_2', \cdots, A_m') = \min D'(A_1, A_2, \cdots, A_m)$,则称 $A_1', A_2', \cdots, A_m'$ 为数据集 A 的一维 m 组聚类 $(m \geq 2)$.

定义5 对数据集 A 中数据按从大到小排序,得到有序组 $B = \{b_1, b_2, \cdots, b_n\}$, $b_i (i \in N)$ 为 A 中第 i 大元素,则称 $\{\Delta_t \mid \Delta_t = b_t - b_{t+1}, t = 1, \cdots, n-1\}$ 为 A 的有序增量集, Δ_t 为 A 的有序增量.

下面给出一种一维聚类的方法 ——有序增量分割法,其有效性证明参见定理1. 算法具体步骤如下:

步骤1: 对于 $A = \{a_1, a_2, \cdots, a_n\}$,按定义5求得 A 的有序增量集 $\{\Delta_t \mid t = 1,2,\cdots,n-1\}$.

步骤2: 给定 $m (2 \leq m \leq n)$,按由小到大的顺序依次选取前 $m-1$ 个最大的有序增量 Δ_t,并在产生 $m-1$ 个 Δ_t 的数据之间进行分割,数轴上独立的 m 个数据群 $A_1', \cdots, A_m'$ 即为要求的一维 m 组聚类.

如果步骤2中出现 Δ_t 相等的情况,则按如下原则处理:

1) 远中原则,即在离中位数较远的位置进行分割. 若仍无法区分,则选择离算术平均值较远的位置进行分割.

2) 近中原则,与远中原则恰好相反.

远中原则体现了决策者偏好中间数据、厌恶风险的思想;近中原则体现了决策者偏好极端数据、追求风险的思想.

引理1 A_1 和 A_2 是数据集 A 的任意一个划分, A_2' 和 A_2' 是数据集 A 的一个划分, B 为对 A 中元素从大到小排列的数据集, $\Delta_t (t = 1, \cdots, n-1)$ 为 A 的有序增量. 使得 $D'(A_2', A_2') = \min D'(A_1, A_2)$ 成立的充分必要条件是:

1) $[\min(A_1'), \max(A_1')] \cap [\min(A_2'), \max(A_2')] = \varnothing$;

2) $\max\{\Delta_t\} = D(A) - D'(A_1', A_2')$.

证明略.

定理1 $A_1, \cdots, A_m$ 为数据集 A 的任意一个划分, $A_1', \cdots, A_m'$ 为 A 用有序增量分割法得到的一维 $m (m \geq 2)$ 组聚类,则 $A_1', \cdots, A_m'$ 满足 $D'(A_1', \cdots, A_m') = \min D'(A_1, \cdots, A_m)$.

证明(采用归纳法) 当 $m = 2$ 时,由有序增量分割法的步骤可知, A_1' 和 A_2' 满足引理1中的条件1) 和2),因而 $D'(A_1', A_2') = \min D'(A_1, A_2)$.

假设 $m = k (k > 2$,即 $k-1$ 次分割),有 $D'(A_1', A_2', \cdots, A_k') = \min D'(A_1, A_2, \cdots, A_k)$ 成立. 在第 k 次分割时,对 $A_i' (i = 1,2,\cdots,k)$ 用有序增量分割法得到的一维两组聚类为 $A_i^{(1)}$ 和 $A_i^{(2)}$, $D'(A_i^{(1)}, A_i^{(2)}) = D(A_i') - \max_{t_i'}\{\Delta_{i,t_i'}\}$ ($\max_{t_i'}\{\Delta_{i,t_i'}\}$ 为 A_i' 中最大的有序增量),由有序增量分割法的步骤2,可知 $\max_{t_i'}\{\Delta_{i,t_i'}\} = \max_i \max_{t_i'}\{\Delta_{i,t_i'}\}$. 设 $A_1, A_2, \cdots, A_k, A_{k+1}$ 为 A 的任意一个划分,则有

$$D'(A_1, A_2, \cdots, A_k, A_{k+1}) =$$

$$D'(A_1, A_2, \cdots, A_k) - \Delta_{i,t_i} \geq$$

$$\min D'(A_1, A_2, \cdots, A_k) - \Delta_{i,t_i} \geq$$

$$D'(A_1', A_2', \cdots, A_k') - \max_i \max_{t_i'}\{\Delta_{i,t_i'}\} =$$

$$D'(A_1', \cdots, A_i^{(1)}, A_i^{(2)}, \cdots, A_k'),$$

其中 t_i 为 A_i 中有序增量的下标.

将 $A_1', \cdots, A_i^{(1)}, A_i^{(2)} \cdots, A_k'$ 改写为 $A_1', A_2', \cdots, A_k', A_{k+1}'$,可得

$$D'(A_1, A_2, \cdots, A_k, A_{k+1}) \geq$$

$$D'(A_1', A_2', \cdots, A_k', A_{k+1}') =$$

$$\min D'(A_1, A_2, \cdots, A_k, A_{k+1}). \square$$

3 密度加权平均中间算子

3.1 算子定义

设 $A_1, \cdots, A_m$ 为 A 进行一维聚类后按元素个数由大到小排序的 m 个数据集合, $A_j (j \in M)$ 中数据元素个数为 $k_j (1 \leq k_j \leq n-1)$,当 $j_1 < j_2$ 时,满足 $k_{j_1} \geq k_{j_2}$,则称 $A_1, \cdots, A_m$ 为序化后 A 的一维 m 组聚类(以下 $A_1, \cdots, A_m$ 均指序化后的聚类结果).

定义6 对于数据集 A,设 DWA: $R^n \to R$,若

$$\mathrm{DWA}_{\Lambda,\xi}(a_1, a_2, \cdots, a_n) = \sum_{j=1}^{m} \xi_j \Lambda(A_j), \quad (1)$$

其中: $A_1, \cdots, A_m$ 为序化后 A 的一维 m 组聚类, $\xi = (\xi_1, \xi_2, \cdots, \xi_m)$ 为一密度加权向量, $\xi_j \in [0,1]$, $\sum_{j=1}^{m} \xi_j$

$=1, j\in M$, Λ 为某一信息集结算子. 则称函数 DWA 为密度加权平均中间算子,也称 DWA 算子.

3.2　密度加权向量及性质

式(1)中的密度加权向量 $\xi=(\xi_1,\xi_2,\cdots,\xi_m)$ 可按下式确定:

$$\xi_j=\beta_j(k_j/n)\Big/\sum_{j=1}^{m}\beta_j(k_j/n), j\in M. \qquad (2)$$

其中 $\beta_j\geq 0$ 为密度影响因子,可设置成线性或非线性的. 这里给出非线性形式的一种配置方式

$$\beta_j=(k_j/n)^{\alpha}, j\in M. \qquad (3)$$

其中:$\alpha\in(-\infty,+\infty)$ 为密度影响指数(一般 α 在 $[-10,10]$ 上取值即可满足需要), $\beta_j\in(0,1)$.

密度权向量有趋同性、趋中性和趋极性之分,不同性质的权向量体现不同的决策规则. 决策者强调主体信息或群体共识,可选择同性权向量;强调极端信息或个别意见,可选择极性权向量;中性权向量是一种过渡性权向量,本身不体现对数据分布的偏好.

定义 7　当 $\xi=(k_1/n,k_2/n,\cdots,k_m/n)$ 时,称 ξ 为中性密度加权向量,记为 ξ_{Ave};当 $\xi=(1,0,\cdots,0)$ 时,称 ξ 为全同性密度加权向量,记为 ξ_{Sam};当 $\xi=(0,0,\cdots,1)$ 时,称 ξ 为全极性密度加权向量,记为 ξ_{Ext}.

性质介于 ξ_{Ave} 与 ξ_{Sam} 之间的 ξ 为偏同性密度加权向量,性质介于 ξ_{Ave} 与 ξ_{Ext} 之间的 ξ 为偏极性密度加权向量.

定义 8　对于任一密度加权向量 ξ,其同性程度的度量为

$$T_s(\xi)=\frac{1}{m-1}\sum_{j=1}^{m}\Big[\frac{(m-j)\xi_j}{k_j/n}\cdot\frac{1}{\sum_{j=1}^{m}\xi_j(n/k_j)}\Big], \qquad (4)$$

其极性程度的度量为

$$T_e(\xi)=1-T_s(\xi). \qquad (5)$$

容易证明有下述结论成立:

1) 对于式(4), $T_s(\xi)\in[0,1]$, $T_s(\xi_{\mathrm{Ave}})=0.5$, $T_s(\xi_{\mathrm{Sam}})=1$, $T_s(\xi_{\mathrm{Ext}})=0$.

2) 对于式(3),当 $\alpha>0$ 时, $T_s(\xi)\in(0.5,1)$;当 $\alpha<0$ 时, $T_s(\xi)\in(0,0.5)$.

定义 9　对于任一密度加权向量 ξ,若 $T_s(\xi)\in(0.5,1)$ 或 $T_e(\xi)\in(0,0.5)$,则称 ξ 为偏同性密度加权向量;若 $T_s(\xi)\in(0,0.5)$ 或 $T_e(\xi)\in(0.5,1)$,则称 ξ 为偏极性密度加权向量.

由式(2)～(4)求得

$$T_s[\xi(\alpha)]=\frac{1}{m-1}\sum_{j=1}^{m}\Big[(m-j)\frac{(k_j/n)^{\alpha}}{\sum_{j=1}^{m}(k_j/n)^{\alpha}}\Big]. \qquad (6)$$

在式(6)中, $T_s[\xi(\alpha)]$ 是 α 的严格递增函数, ξ 由 α 唯一确定,因而 $T_s[\xi(\alpha)]$ 与 ξ 一一对应.

基于上述分析,决策者可按如下两种方式确定密度权向量 ξ:

1) 满意值法. 初始时任选一 α 值,根据式(6)求得 $T_s[\xi(\alpha)]$,若 $T_s[\xi(\alpha)]$ 值满意,则将 α 代入式(2)和(3)求得 ξ;否则重新选择 α,重复上述过程,直到 $T_s[\xi(\alpha)]$ 值满意为止.

2) 目标值法. 决策者提供反映偏好水平的 $T_s[\xi(\alpha)]$ 值,由式(6)经数值迭代计算得到 α 值,将 α 代入式(2)和(3)求得 ξ.

4　基于 DWA 的 5 种合成算子

DWA 为中间算子,需要与其他集结算子结合使用. 以下给出 DWA 与 WAA[6,11],OWA[6,10],AA[6,7],Min[6,7,12],Max[6,7,12] 结合使用的 5 种合成算子.

定义 10　设 $\mathrm{DWA_{WAA}}: R^n\rightarrow R$,若

$$\mathrm{DWA}_{\mathrm{WAA},w,\xi}(a_1,a_2,\cdots,a_n)=\sum_{i=1}^{m}\xi_i\mathrm{WAA}(A_i)=\sum_{i=1}^{m}\xi_i\Big[\sum_{j=1}^{k_i}w_j^{(i)}b_j^{(i)}\Big], \qquad (7)$$

其中:$A_i=\{b_j^{(i)}\mid i\in M; j=1,2,\cdots,k_i\}$, $\sum_{i=1}^{m}k_i=n$, $b_j^{(i)}$ 为数据集 A 中一数据元素, $\xi=(\xi_1,\xi_2,\cdots,\xi_m)$ 为一密度加权向量, $w_i=(w_1^{(i)},w_2^{(i)},\cdots,w_{k_i}^{(i)})$ 为 $A_i=(b_1^{(i)},b_2^{(i)},\cdots,b_{k_i}^{(i)})$ 中元素重要性的归一化加权向量,满足 $\sum_{j=1}^{k_i}w_j^{(i)}=1$, $w_j^{(i)}>0$. 则称 $\mathrm{DWA_{WAA}}$ 为密度算术加权平均算子,也称 $\mathrm{DWA_{WAA}}$ 算子.

定义 11　设 $\mathrm{DWA_{OWA}}: R^n\rightarrow R$,若

$$\mathrm{DWA}_{\mathrm{OWA},\omega,\xi}(a_1,a_2,\cdots,a_n)=\sum_{i=1}^{m}\xi_i\mathrm{OWA}(A_i)=\sum_{i=1}^{m}\xi_i\Big[\sum_{j=1}^{k_i}\omega_j^{(i)}b_j^{(i)}\Big], \qquad (8)$$

其中:$A_i=\{b_j^{(i)}\mid i\in M; j=1,2,\cdots,k_i\}$, $\sum_{i=1}^{m}k_i=n$, $b_j^{(i)}$ 为 A_i 中第 j 大元素,且为数据集 A 中一数据元素, $\xi=(\xi_1,\xi_2,\cdots,\xi_m)$ 为一密度加权向量, $\omega_i=(\omega_1^{(i)},\omega_2^{(i)},\cdots,\omega_{k_i}^{(i)})$ 为 $A_i=(b_1^{(i)},b_2^{(i)},\cdots,b_{k_i}^{(i)})$ 中元素在 A 中位置重要性的归一化加权向量,满足 $\sum_{j=1}^{k_i}\omega_j^{(i)}=1$, $\omega_j^{(i)}\geq 0$. 则称 $\mathrm{DWA_{OWA}}$ 为密度有序加权平均算子,也称 $\mathrm{DWA_{OWA}}$ 算子.

定义 12　设 $\mathrm{DWA_{AA}}: R^n\rightarrow R$,若

$$\mathrm{DWA_{AA}}(a_1,a_2,\cdots,a_n)=\sum_{i=1}^{m}\xi_i\mathrm{AA}(A_i)=\sum_{i=1}^{m}\xi_i\Big[\frac{1}{k_i}\sum_{j=1}^{k_i}b_j^{(i)}\Big], \qquad (9)$$

则称 $\mathrm{DWA_{AA}}$ 为密度算术平均算子.

定义 13　设 $\mathrm{DWA_{Min}}: R^n\rightarrow R$,若

$$\mathrm{DWA}_{\mathrm{Min},\xi}(a_1, a_2, \cdots, a_n) = \sum_{i=1}^{m} \xi_i \min(A_i), \tag{10}$$

则称 $\mathrm{DWA}_{\mathrm{Min}}$ 为密度最小算子,也称密度“与”算子.

定义 14 设 $\mathrm{DWA}_{\mathrm{Max}}: R^n \to R$,若

$$\mathrm{DWA}_{\mathrm{Max},\xi}(a_1, a_2, \cdots, a_n) = \sum_{i=1}^{m} \xi_i \max(A_i), \tag{11}$$

则称 $\mathrm{DWA}_{\mathrm{Max}}$ 为密度最大算子,也称密度“或”算子.

定义 12 ～ 14 中字符的含义与定义 10 中相同.

$\mathrm{DWA}_{\mathrm{WAA}}$ 算子适用于属性权重已知的属性值信息集结,或决策者权重已知的决策群体信息集结; $\mathrm{DWA}_{\mathrm{OWA}}$, $\mathrm{DWA}_{\mathrm{AA}}$, $\mathrm{DWA}_{\mathrm{Min}}$, $\mathrm{DWA}_{\mathrm{Max}}$ 算子适用于属性权重未知的属性值信息集结,或决策者权重未知的决策群体信息集结. 合成算子完成对一组数据的二次集结, 因而兼顾了各分算子的特点: $\mathrm{DWA}_{\mathrm{WAA}}$ (或 $\mathrm{DWA}_{\mathrm{OWA}}$) 算子中既考虑了数据本身重要性(或数据值大小位置的重要性),又考虑了数据疏密分布的重要性; $\mathrm{DWA}_{\mathrm{AA}}$ 算子是基于数据疏密分布重要性的算术平均值; $\mathrm{DWA}_{\mathrm{Min}}$ (或 $\mathrm{DWA}_{\mathrm{Max}}$) 算子利用一组数据中多个数据群的最小值(或最大值),相对于 Min(或 Max) 算子而言,密度最大(或最小) 算子的集结值更加稳健.

5 算例分析

以下通过一个数值例子来说明 DWA 算子的特点,并给出一些结论.

设有两个决策方案 x_1 和 x_2, $A_{x_1} = (0.8, 1.2, 4.7, 7.2, 7.6, 8.3, 8.9)$, $A_{x_2} = (10.257, 9.857, 6.357, 3.857, 3.457, 2.757, 2.157)$. A_{x_1} 和 A_{x_2} 中的数据可看成 x_1 和 x_2 的 7 个属性值,也可看成 7 个专家对 x_1 和 x_2 的评价值. 对 x_1 和 x_2 的优劣排序既可理解为单人多准则的综合评价问题,也可理解为多人单准则的群决策问题.

用有序增量分割法对 7 个数据进行一维聚类(分成 3 组),并求解密度权向量 ξ_{x_1} 和 ξ_{x_2} 的算式.

1) 对 A_{x_1} 中元素按从大到小排列,排列后的数据集为 $B_{x_1} = (8.9, 8.3, 7.6, 7.2, 4.7, 1.2, 0.8)$.

2) 求 B_{x_1} 的有序增量集 $\{\Delta_t \mid t = 1, 2, \cdots, 6\}$,用有序增量集中最大的两个元素将 B_{x_1} 分割成 3 组

$$(8.900\underbrace{,}_{\Delta_1=0.6}8.300\underbrace{,}_{\Delta_2=0.7}7.600\underbrace{,}_{\Delta_3=0.4}7.200\overset{\Downarrow_2}{\underbrace{,}_{\Delta_4=2.5}}4.700\overset{\Downarrow_1}{\underbrace{,}_{\Delta_5=3.5}}1.200\underbrace{,}_{\Delta_6=0.4}0.800),$$

$\Downarrow_1$ 和 $\Downarrow_2$ 分别标示了第 1 次和第 2 次分割的位置. 分割后的 3 组数据分别为 $A_{x_1}^{(1)} = (8.9, 8.3, 7.6, 7.2)$, $A_{x_1}^{(2)} = (1.2, 0.8)$, $A_{x_1}^{(3)} = (4.7)$. 可见 A_{x_1} 中数据的分布是向大数偏斜的.

3) 设 $\xi_{x_1} = (\xi_{x_1}^{(1)}, \xi_{x_1}^{(2)}, \xi_{x_1}^{(3)})$,由式(2) 和(3) 可知

$$\xi_{x_1}^{(i)} = \eta_{x_1}^{(i)} / \sum_{i=1}^{3} \eta_{x_1}^{(i)}, i = 1, 2, 3. \tag{12}$$

其中

$$\eta_{x_1}^{(1)} = (4/7)^{\alpha+1}, \eta_{x_1}^{(2)} = (2/7)^{\alpha+1}, \eta_{x_1}^{(3)} = (1/7)^{\alpha+1}.$$

同理,可得 A_{x_2} 分割后的 3 组数据分别为 $A_{x_2}^{(1)} = (3.857, 3.457, 2.757, 2.157)$, $A_{x_2}^{(2)} = (10.257, 9.857)$, $A_{x_2}^{(3)} = (6.357)$. 可见 A_{x_2} 中数据的分布是向小数偏斜的. 相应的密度权向量 $\xi_{x_2} = (\xi_{x_2}^{(1)}, \xi_{x_2}^{(2)}, \xi_{x_2}^{(3)})$ 为

$$\xi_{x_2}^{(i)} = \eta_{x_2}^{(i)} / \sum_{i=1}^{3} \eta_{x_2}^{(i)}, i = 1, 2, 3. \tag{13}$$

表 1 x_1 方案下 7 种算子随 α 的取值情况

α	$T_s[\xi(\alpha)]$	$\mathrm{DWA}_{\mathrm{AA}}$	WAA	$\mathrm{DWA}_{\mathrm{WAA}}$	OWA	$\mathrm{DWA}_{\mathrm{OWA}}$	$\mathrm{DWA}_{\mathrm{Min}}$	$\mathrm{DWA}_{\mathrm{Max}}$
- 8	0.002 0	4.671 5	6.191 5	4.672 0	5.560 0	4.673 0	4.669 9	4.6731
- 4	0.033 0	4.339 7	6.191 5	4.347 5	5.560 0	4.353 3	4.306 8	4.374 0
- 2	0.142 9	4.114 3	6.191 5	4.141 1	5.560 0	4.084 2	3.942 9	4.300 0
- 1	0.285 7	4.566 7	6.191 5	4.608 0	5.560 0	4.429 8	4.233 3	4.933 3
- 0.5	0.386 7	5.009 8	6.191 5	5.057 5	5.560 0	4.797 2	4.583 2	5.481 6
0	0.500 0	5.528 6	6.191 5	5.581 2	5.560 0	5.236 8	5.014 3	6.100 0
0.5	0.613 3	6.047 2	6.191 5	6.103 2	5.560 0	5.682 1	5.458 3	6.703 7
1	0.714 3	6.509 5	6.191 5	6.567 6	5.560 0	6.082 5	5.861 9	7.233 3
2	0.857 1	7.187 7	6.191 5	7.247 5	5.560 0	6.674 3	6.464 4	7.998 6
4	0.967 0	7.785 0	6.191 5	7.845 3	5.560 0	7.199 5	7.003 9	8.662 9
8	0.998 0	7.986 3	6.191 5	8.046 6	5.560 0	7.377 4	7.187 5	8.885 0

注:OWA 算子的权重为在“大部分”准则[1] 下得到的,为(0, 0.171 4, 0.285 7, 0.285 7, 0.257 1, 0, 0); WAA 算子的权重随机生成,为(0.052 7, 0.102 1, 0.233 9, 0.128 0, 0.202 6, 0.053 9, 0.226 8).

表 2　x_2 方案下 7 种算子随 α 的取值情况

α	$T_s[\xi(\alpha)]$	DWA_{AA}	WAA	DWA_{WAA}	OWA	DWA_{OWA}	DWA_{Min}	DWA_{Max}
- 8	0.002 0	6.385 6	4.865 6	6.385 1	4.197 1	6.385 6	6.384 0	6.387 2
- 4	0.033 0	6.717 4	4.865 6	6.709 6	4.197 1	6.722 8	6.683 2	6.750 3
- 2	0.142 9	6.942 9	4.865 6	6.916 0	4.197 1	6.998 9	6.757 1	7.114 3
- 1	0.285 7	6.490 5	4.865 6	6.449 1	4.197 1	6.621 2	6.123 8	6.823 8
- 0.5	0.386 7	6.047 4	4.865 6	5.999 6	4.197 1	6.225 1	5.575 5	6.473 9
0	0.500 0	5.528 6	4.865 6	5.475 9	4.197 1	5.752 7	4.957 1	6.042 9
0.5	0.613 3	5.010 0	4.865 6	4.954 0	4.197 1	5.275 3	4.353 5	5.598 9
1	0.714 3	4.547 6	4.865 6	4.489 5	4.197 1	4.846 5	3.823 8	5.195 2
2	0.857 1	3.869 5	4.865 6	3.809 6	4.197 1	4.213 4	3.058 5	4.592 8
4	0.967 0	3.272 2	4.865 6	3.211 8	4.197 1	3.652 2	2.394 2	4.053 3
8	0.998 0	3.070 8	4.865 6	3.010 5	4.197 1	3.462 3	2.172 2	3.869 6

注:OWA 和 WAA 算子的权重设置同表 1.

其中

$\eta_{x_2}^{(1)} = (4/7)^{\alpha+1}$, $\eta_{x_2}^{(2)} = (2/7)^{\alpha+1}$, $\eta_{x_2}^{(3)} = (1/7)^{\alpha+1}$.

由式(1) ~ (13) 可得到方案 x_1 和 x_2 在不同密度影响指数 α 值下,WAA 和 OWA 以及 5 种合成算子的集结值,如表 1 和表 2 所示(仅列出 11 组离散情形的取值).

分析表 1 和表 2 中的数据,可得到以下结论:

1) 当 $\alpha = 0$ 时,密度加权向量 ξ 是中性的,此时 $DWA_{AA}(A_{x_1}) = DWA_{AA}(A_{x_2}) = AA(A_{x_1}) = AA(A_{x_2}) = 5.528\ 6$;当 $\alpha < 0$ 时,密度加权向量 ξ 是偏极性的,体现了决策者突出个别意见的决策准则,因而在 A_{x_1} 和 A_{x_2} 中分配给 $A_{x_1}^{(3)}$ 和 $A_{x_2}^{(3)}$ 的权重较大,分配给 $A_{x_1}^{(1)}$ 和 $A_{x_2}^{(1)}$ 的权重较小,故有 $DWA_{AA}(A_{x_1}) < AA(A_{x_1})$, $DWA_{AA}(A_{x_2}) > AA(A_{x_2})$;当 $\alpha > 0$ 时,密度加权向量 ξ 是偏同性的,体现了决策者强调群体共识、弱化分歧意见的决策准则,与 $\alpha < 0$ 时的情形相反, 故有 $DWA_{AA}(A_{x_1}) > AA(A_{x_1})$, $DWA_{AA}(A_{x_2}) < AA(A_{x_2})$.

2) 随着 α 的增大,对于方案 x_1,5 种合成算子的取值是先降后增的;对于方案 x_2,5 种合成算子的取值是先增后降的. 说明随密度权向量由偏极性向偏同性转变,A_{x_1} 和 A_{x_2} 中 3 个数据分组(可看成 3 个意见群) 的作用在不断轮换. 集结值呈现这种规律性变化正好印证了 DWA 算子特殊的集结特征.

3) 在本例中, $T_s(-\infty) = \lim\limits_{\alpha\to-\infty} T_s[\xi(\alpha)] = 0$, $T_s(+\infty) = \lim\limits_{\alpha\to+\infty} T_s[\xi(\alpha)] = 1$, $T_s[\xi(\alpha)]$ 随 α 的增大而增大,并且 $T_s(\alpha) + T_s(-\alpha) = 1$,说明本文给出的非线性赋权方式具有较好的数值特性.

4) DWA_{AA}, DWA_{WAA}, DWA_{OWA} 是介于 DWA_{Min} 与 DWA_{Max} 之间的合成算子. 对于任一数据集 A,满足 $DWA_{AA}(A)$, $DWA_{WAA}(A)$, $DWA_{OWA}(A) \in [DWA_{Min}(A), DWA_{Max}(A)]$. 由此可见,$DWA_{Min}$ 是一个悲观的集结算子,DWA_{Max} 是一个乐观的集结算子.

5) 当 α 取不同值时,采用 5 种合成算子得到的集结值的大小不断变化,方案 x_1 和 x_2 的优劣顺序随之更换. 可见,密度权向量的选择(即决策者的偏好) 在很大程度上影响了方案的排序结论.

6　结　　语

本文提出的 DWA 算子,主要用于多属性群决策中群体决策意见的集结或多个属性值的信息集结. 由于 DWA 算子是基于数据疏密分布信息构造的,从统计意义上说,数据个数越多,算子的集结效果越好. 进一步研究可在 DWA 算子向多种数据形式推广、密度权向量的确定方法及与其他算子的结合上展开.

参考文献(References)

[1] Yager R R. On ordered weighted averaging aggregation operators in multicriteria decisionmaking [J]. IEEE Trans on Systems, Man and Cybernetics, 1988, 18(1): 183-190.

[2] Yager R R, Kacprzyk J. The ordered weighted averaging operators: Theory and applications [M]. Norwell: Kluwer, 1997.

[3] Yager R R. Applications and extensions of OWA aggregation[J]. Int J of Man-machine Study, 2002, 37 (1): 103-132.

[4] Filev D, Yager R R. On the issue of obtaining OWA operator weights[J]. Fuzzy Sets and Systems, 1998, 94 (2): 157-169.

[5] Xu Z S. An overview of methods for determining OWA weights[J]. Int J of Intelligent Systems, 2005, 20(8): 843-865.

(下转第 524 页)

平均每个事务 23 项,文件大小 558 KB), T100120D (870 项,100 000 个事务,平均每个事务 10 项,文件大小 998 KB). 实验结果如图 1～图 4 所示.

从实验结果可以发现,对于数据集 Pumsb 和 Connect-4, MFIM 算法的性能好于 FPMAX*; 对于数据集 Mushroom 和 T100120D, MFIM 算法的性能不如 FPMAX*. 通过对数据集的特征研究发现,前两个数据集属于浓密数据集,后两个数据集属于稀疏数据集. 浓密数据集的 FP-tree 中包含较多的完全前缀路径,本文算法无需对完全前缀路径逐项建立子 FP-tree,减少了冗余操作,因此算法性能有了较大的提高. 本文算法对于 Pumsb 数据集的效果要好于对于 Connect-4 数据集的效果. Pumsb 中的模式属于长模式,其中的完全前缀路径的长度较长,算法减少的冗余操作较多,因此性能提高幅度较大. 对于稀疏数据集,其中没有或少有完全前缀路径,因此算法优化的幅度不大. 因为 FPMAX* 利用数组降低了建立 FP-tree 的代价,所以 FPMAX* 的性能要好于 MFIM 的性能. 通过以上分析可知,本文算法对于浓密数据集的长模式挖掘效果较好.

6 结 语

通过对基于 FP-tree 的最大频繁项集挖掘算法的研究发现,该类算法中有一些不必要的冗余操作. 为进一步对该类算法进行优化,本文提出了完全前缀路径和有序 FP-tree 的概念,给出了建立有序 FP-tree 的算法. 在此基础上,提出了用有序 FP-tree 进行最大频繁项集挖掘的算法 ——MFIM 算法. 该算法利用有序 FP-tree 中的完全前缀路径包含右边子路径这一特点,对挖掘算法进行优化. 实验结果表明,该算法对于浓密数据集中长模式的挖掘具有较大的应用价值. 下一步工作将结合实际应用,对算法进一步改进和完善.

参考文献(References)

[1] Agrawal R, Imielinski T, Swami R. Mining association rules between sets of items in large databases[C]. SIGMOD '93. Washington, 1993: 207-216.

[2] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]. VLDB '94. Santiago, 1994: 487-499.

[3] Grahne G, Zhu J F. Efficiently using prefix-trees in mining frequent itemsets[C]. FIMI '03. Melbourne, 2003.

[4] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[C]. Proc of ACM SIGMOD '00. Dallas, 2000: 1-12.

[5] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation: A frequent-pattern tree approach[J]. Data Mining and Knowledge Discovery, 2004, 8(1): 53-87.

[6] He Z, Deng S, Xu X. A FP-tree based approach for mining all strongly correlated pairs[C]. Proc of Conf on Computational Intelligence and Security. Xi'an, 2005: 735-740.

[7] Han J, Wang J, Lu Y, et al. Mining top-*K* frequent closed patterns without minimum support[C]. ICDM '02. Washington, 2003: 211-218.

[8] Zaki M J, Hsiao C. CHARM: An efficient algorithm for closed itemset mining[C]. SDM '02. Arlington, 2002: 457-473.

[9] Xin D, Han J W, Yan X F, et al. Mining compressed frequent-pattern sets[C]. Proc of the 31st VLDB Conf. Trondheim, 2005: 709-720.

[10] Grahne G, Zhu J. High performance mining of maximal frequent itemsets[C]. Proc of the 6th SIAM Int'l Workshop on High Performance Data Mining. San Francisco, 2003: 135-143.

(上接第 519 页)

[6] Xu Z S, Da Q L. An overview of operators for aggregating information[J]. Int J of Intelligent Systems, 2003, 18(9): 953-969.

[7] Yager R R. Families of OWA operators[J]. Fuzzy Sets and Systems, 1993, 59(1): 125-148.

[8] Herrera F, Herrera Viedma E, Verdegay J L. Direct approach processes in group decision making using linguistic OWA operators[J]. Fuzzy Sets and Systems, 1996, 79(2): 175-190.

[9] Yager R R. Induced ordered weighted averaging operators[J]. IEEE Trans on Systems, Man and Cybernetics: Part B, 1999, 29(2): 141-150.

[10] Xu Z S, Da Q L. The uncertain OWA operator[J]. Int J of Intelligent Systems, 2002, 17(6): 569-575.

[11] Keeney R L, Raiffa H. Decisions with multiple objectives: Preferences and value trade-offs[M]. New York: Wiley, 1976.

[12] Dubois D, Prade H. A review of fuzzy sets aggregation connectives[J]. Information Science, 1985, 36(1): 85-121.