

FCMBP 模糊聚类算法的改进

华 斌, 张洪波, 何 晓

(天津财经大学理工学院, 天津 300222)

摘 要: 在 FCMBP 算法中, 高阶模糊等价标准型的平移等价类数据库缺少一个高效的生成算法, 且每一个模糊等价标准型的平移等价类需要定义相应的相似参数系等价类, 过程繁琐。为解决上述问题, 提出由低阶向高阶自动生成模糊等价标准型矩阵的平移等价类数据库的算法以及生成相应相似参数系的等价类的算法。通过实例验证该算法较好地解决了高阶模糊等价标准型的平移等价类数据库的自动生成问题。

关键词: 知识发现; 模糊聚类; FCMBP 算法

Improvement of FCMBP Fuzzy Clustering Algorithm

HUA Bin, ZHANG Hong-bo, HE Xiao

(Institute of Technology, Tianjin University of Finance & Economics, Tianjin 300222, China)

【Abstract】 In FCMBP algorithm, there is no effective generation algorithm in the database of the translation equivalence classes of high-order fuzzy equivalent standard matrix in the past time, and it is very complicated to define the corresponding database of similar parameters system for each translation equivalence class of fuzzy equivalent standard matrix. In order to solve the problems, this paper proposes an automatic generation algorithm for the database of the translation equivalence classes with the low to high standard fuzzy equivalent matrix, and an algorithm for generating the corresponding database of the equivalent classes of similar parameters system. The case proves that the algorithm effectively solve the problem of automatic generation for the database of the translation equivalence classes of high-order fuzzy equivalent standard matrix.

【Key words】 knowledge discovery; fuzzy clustering; FCMBP algorithm

DOI: 10.3969/j.issn.1000-3428.2011.02.065

1 概述

模糊聚类是一类获得广泛应用的数据挖掘方法。文献[1]提出一种新的模糊聚类算法 NSFCM, 将其应用于文本挖掘中。文献[2]基于模糊聚类研究了快速路 VMS 信息发布方法。

由于模糊相似矩阵一般不具有传递性, 因此模糊聚类一般不使用模糊相似矩阵直接聚类。常用的模糊聚类方法是先通过标定得到模糊相似矩阵 R , 再求 R 的传递闭包 $t(R)$, 从而将 R 改造成一个模糊等价矩阵 R^* , 最后根据 R^* 求聚类图。然而求传递闭包的过程经过了一系列的非恒等变换, 使得 R^* 与 R 在形式上往往有较大的差别, 因此, 由 R^* 得到的聚类是否真实地反映了原始问题的聚类情况, 这在理论上无法得到严格的保证^[3], 即这种方法的“失真”问题没有得到根本的解决。文献[4]给出了一种基于摄动解决“失真”问题的模糊聚类方法 FCMBP, 其基本思想是: 试图寻找一个与原始模糊相似矩阵 R 按某种“距离”最小的模糊等价矩阵 R^0 , 再根据 R^0 进行聚类。文献[5-6]证明了失真最小的模糊等价阵, 即 Fuzzy 最优等价阵的存在性, 为 Fuzzy 聚类提供了理论依据。文献[7]从理论上证明了 FCMBP 模糊聚类方法比传递闭包法失真小, 而且在基于模糊相似矩阵的模糊聚类方法中 FCMBP 方法的失真最小, 同时举例说明了 FCMBP 方法不仅失真最小, 而且有时与传递闭包法的聚类结果有本质差异。

本文基于 FCMBP 的理论研究了由低阶向高阶生成模糊等价标准型矩阵的平移等价类算法, 并通过定义初始低阶数据库, 利用函数自身嵌套调用的方法, 解决了高阶数据库的生成问题, 给出了改进运算速度的多阶段逐步改变函数初始条件的思路。另外, 针对本文给出的模糊等价标准型平移等价类的数据库生成算法, 给出了一个生成相应相似参数系的

等价类算法, 解决了为每一个模糊等价标准型平移等价类一配备相应相似参数系等价类的问题。

2 FCMBP 模糊聚类的理论基础和经典算法

2.1 FCMBP 模糊聚类的理论基础

记 Y_n 为全体 n 阶 Fuzzy 相似矩阵的集合, X_n 为全体 n 阶 Fuzzy 等价矩阵的集合, 即 n 阶 Fuzzy 相似矩阵方程 $X^2=X$ 的解集, S_n 为 n 元置换群, J 为全体参数系集合。

定理 1 若 $X \in Y_n, X = (x_{ij})_{n \times n}$, 则 $X \in X_n \Leftrightarrow x_{ik} \wedge x_{kj} = x_{ij}, \forall i \neq j \neq k \quad n$ 。

定理 2 设 $X = (x_{ij})_{n \times n}, \sigma \in S_n$, 若 $X \in Y_n$, 则 $X_\sigma = (x_{\sigma(i)\sigma(j)})_{n \times n} \in Y_n$, 若 $X \in X_n$, 则 $X_\sigma = (x_{\sigma(i)\sigma(j)})_{n \times n} \in X_n$ 。

定理 3 设 $X \in Y_n$, 若 $\exists t \in [0, 1]$ 和 $\{1, 2, \dots, n\}$ 的一个不交分解 $I_m \cup I_m^c$, 使得:

$$X(I_m) \in X_n, X(I_m^c) \in X_{n-m} \quad (1)$$

$$X(I_m) = (t)_{m \times m}, X(I_m^c) = (t)_{(n-m) \times (n-m)} \quad (2)$$

$$X(I_m, I_m^c) = (t)_{m \times (n-m)}, X(I_m^c, I_m) = (t)_{(n-m) \times m} \quad (3)$$

则 $X \in X_n$, 其中,

$$X(I_m) = (x_{ij})_{m \times m}, i \in I_m, j \in I_m$$

$$X(I_m^c) = (x_{ij})_{(n-m) \times (n-m)}, i \in I_m^c, j \in I_m^c$$

$$X(I_m^c, I_m) = (x_{ij})_{(n-m) \times m}, i \in I_m^c, j \in I_m$$

基金项目: 国家科技部创新基金资助项目(09c26211200200)

作者简介: 华 斌(1963 -), 男, 教授, 主研方向: 知识发现, 人工智能; 张洪波、何 晓, 硕士研究生

收稿日期: 2010-06-15 **E-mail:** greenrey2003@yahoo.com.cn

$$X(I_m, I_m^c) = (x_{ij})_{m \times (n-m)}, i \in I_m, j \in I_m^c$$

定义1 设 $X \in Y_n$, 若 X 存在一个满足式(1)~式(3)的分解, 则称 X 有一个分解构造, 记为 $(X(I_m), X(I_m^c), X(I_m, I_m^c), X(I_m^c, I_m))$ 。

定理4 如果 $X \in Y_n$, 则 $X \in X_n \Leftrightarrow X$ 有一个分解构造。

对于 $\forall X \in X_n$, 得到 X 的第1次分解构造后, 因为 $X(I_m) \in X_n, X(I_m^c) \in X_{n-m}$, 所以对 $X(I_m)$ 和 $X(I_m^c)$ 仍可做分解构造。如此分解下去直到子阵和余阵的阶为1为止。本文把每次分解所依据的参数 t 提取出来构成反映分解过程的图。图1给出了第1次的分解过程。

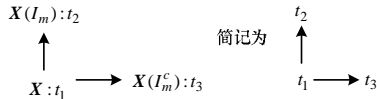


图1 分解构造过程

定义2 设 $X \in Y_n$, 当把 X 完全分解时, 所有参数所构成的反映分解过程的图称为 X 的参数系。

定理5 $\forall X \in X_n$, 如果 $(X(I_m), X(I_m^c), X(I_m, I_m^c), X(I_m^c, I_m))$ 是 X 的一个分解构造, 则存在 $\sigma \in S_n$, 使得:

$$X_\sigma = \begin{bmatrix} X(I_m) & X(I_m, I_m^c) \\ X(I_m^c, I_m) & X(I_m^c) \end{bmatrix}$$

定义3(标准分解构造) 对 $\forall X \in X_n$ 进行下列分解:

(1)取 $t = \wedge\{x_{ij}: 1 \leq i \neq j \leq n\}$ 。

(2)若 $t=1$, 则 $X = (1)_{n \times n}$, 此时取 $I_m = \{1\}$, $I_m^c = \{2, 3, \dots, n\}$ 。若 $t < 1$, 在 X 中找出含 t 最多的首列, 不妨设为第 j_0 列。设其中所有等于 t 的元素为: $x_{i_1 j_0} = x_{i_2 j_0} = \dots = x_{i_{n-m} j_0} = t$ ($i_1 < i_2 < \dots < i_{n-m}$)。令 $I_m^c = \{i_1, i_2, \dots, i_{n-m}\}$, $I_m = \{1, 2, \dots, n\} \setminus I_m^c$ 。

(3)取 $\sigma = (i_1, m+1)(i_2, m+2) \dots (i_{n-m}, n)$, 则:

$$X_\sigma = \begin{bmatrix} X(I_m) & X(I_m, I_m^c) \\ X(I_m^c, I_m) & X(I_m^c) \end{bmatrix}$$

本文称这一分解过程为标准分解过程, 称

$$\begin{bmatrix} X(I_m) & X(I_m, I_m^c) \\ X(I_m^c, I_m) & X(I_m^c) \end{bmatrix} \text{ 为 } X \text{ 的标准分解构造。}$$

定义4(等价标准型) 称 $\forall X \in X_n$ 为一个 Fuzzy 置换等价标准型, 简称为等价标准型, 如果 X 的下三角阵有以下分块形式:

(1)设 X_0 是 X 的分块阵的任一子阵, 则 $X_0 = (t)_{m_0 \times n_0}$ 且 $m_0 \leq n_0$, 设 $x_{i_0 j_0}$ 为 X_0 的右上角元素, 则 $i_0^* = j_0^* + 1$, 即右上角元素以1为邻。

(2)设 X_1, X_2 为 X 的分块阵的2个子阵, $X_1 = (t_1)_{m_1 \times n_1}$, $X_2 = (t_2)_{m_2 \times n_2}$, X_2 位于 X_1 上方, 则 $t_1 < t_2$, 即位于较上块的元素大于位于较下块的元素。

(3)设 X_1, X_2 为 X 的分块阵的2个子阵: $X_1 = (t_1)_{m_1 \times n_1}$, $X_2 = (t_2)_{m_2 \times n_2}$, X_2 位于 X_1 右方, 则 $t_1 < t_2$ 或者 $t_1 = t_2, n_1 \leq n_2$, 即位于较右块的元素大于等于位于较左块的元素, 每块的行数大于等于列数。

(4) X 的下三角被分成 $n-1$ 块。

定义5 $\forall X \in X_n$, 若 $\exists \sigma \in S_n$ 使得 X_σ 是一个等价标准型, 则称 X_σ 为 X 的等价标准型。

定理6 $\forall X, Y \in X_n$, 若 $\exists \sigma \in S_n$ 使得 $X_\sigma = Y$, 记为 $X \sim Y$,

则有: $X \sim Y \Leftrightarrow X$ 与 Y 有相同的等价标准型 $\Leftrightarrow X$ 与 Y 有相同的标准参数系。

定理7 设 X 为一个等价标准型, 对其按标准分解过程分块, 若 $X_0 = (t_0)_{m_0 \times n_0}$ 为 X 的下三角阵中任一子阵, 则得出以下结论:

(1)若 X_1, X_2, \dots, X_s 为 X_0 的所有右方相邻子阵, 其行数分别为 m_1, m_2, \dots, m_s , 则 $m_0 = m_1 + m_2 + \dots + m_s + 1$, 特别地, 当 X_0 无右方相邻子阵时, $m_0 = 1$ 。

(2)若 X_1, X_2, \dots, X_s 为 X_0 的所有上方相邻子阵, 其列数分别为 n_1, n_2, \dots, n_s , 则 $n_0 = n_1 + n_2 + \dots + n_s + 1$, 特别地, 当 X_0 无上方相邻子阵时, $n_0 = 1$ 。

在 X_n 和全体参数系集合 J 中引入另一个关系 “ \approx ”:
 $X \approx Y \Leftrightarrow X$ 和 Y 有图结构相同但数值未必相同的参数系, 这是等价关系, 称 X 与 Y 平移等价, 称其相应的参数系为相似参数系。所有等价标准型记为 X_n , 显然 “ \approx ” 在 X_n 也是等价关系, 所以, 得到平移等价标准型的商集 X_n / \approx 。记 J / \approx 为相似参数系的等价类全体, $X(S)$ 为参数系 S 给出的模糊等价标准型, 则有以下结论:

定理8 $X_n = \bigcup_{\sigma \in S_n} \bigcup_{T \in J / \approx} \bigcup_{S \in \tilde{T}} \{X(S)_\sigma\}$, 记 $C(\tilde{T}) = \{S: S \text{ 和 } T \text{ 有相同的参数图, 但序关系是非严格的}\}$, $X = X(T)$, $C(\tilde{X}) = \{X(S): S \in C(\tilde{T})\}$ 为 $C(\tilde{T})$ 对应的模糊等价阵的全体, 则有:
 $X_n = \bigcup_{\sigma \in S_n} \bigcup_{T \in J / \approx} \bigcup_{S \in C(\tilde{T})} \{X(S)_\sigma\}$ 。记距离公式为:

$$d(A, B) = \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (a_{ij} - b_{ij})^2}, A, B \in Y_n$$

定理9 记 $\tilde{k}(n) = |X_n / \approx|$, 则:

$$\tilde{k}(1) = 1$$

$$\tilde{k}(n) = \sum_{m=1}^{n-1} \tilde{k}(m) \tilde{k}(n-m)$$

定理10 对于 $\forall R \in Y_n$, $\exists R^\# \in X_n$, 使得:

$$d(R^\#, R) = \inf_{X \in X_n} d(X, R)$$

本文称与给定的模糊相似矩阵距离最近的模糊等价矩阵为最优模糊等价矩阵。显然最优模糊等价矩阵是相对于给定的模糊相似矩阵失真最小的模糊等价矩阵, 因此, 用它聚类更合理。

定理11 假设 $R \in Y_n$, $\exists R^\# \in C(\tilde{X}) \subseteq X_n$, 使得:

$$d(R^\#, R) = \inf_{X \in C(\tilde{X})} d(X, R)$$

那么:

$$t_i = \frac{b_{i1} + b_{i2} + \dots + b_{im_i}}{m_i} \quad (i = 1, 2, \dots, n-1)$$

其中, t_1, t_2, \dots, t_{n-1} 是 $R^\#$ 的参数; $b_{i1}, b_{i2}, \dots, b_{im_i}$ 是 R 中对应于 $R^\#$ 中 t_i ($i = 1, 2, \dots, n-1$) 的元素。

2.2 FCMBP 模糊聚类的经典算法

由以上理论基础可以得出如下最优模糊等价矩阵的经典算法:

Step1 建立模糊等价标准型的平移等价类数据库 X_n / \approx 以及相似参数系的等价类数据库 J / \approx 。

Step2 取 $R \in Y_n$ 。

Step3 取 $\sigma \in S_n$ 。

Step4 取 $\tilde{X} \in X^- / \approx$ 。

Step5 计算 R_σ 。

Step6 找出 R_σ 中对应于 \tilde{X} 中 t_i 的 $b_{i1}, b_{i2}, \dots, b_{im_i}$ ($i=1, 2, \dots, n-1$)。

Step7 计算 $\hat{t}_i = \frac{b_{i1} + b_{i2} + \dots + b_{im_i}}{m_i}$ ($i=1, 2, \dots, n-1$)。

Step8 检验 \hat{t}_i ($i=1, 2, \dots, n-1$) 是否满足 \tilde{X} 给定的不等式, 若满足, 则转下一步, 否则, 转 Step4。

Step9 利用 \hat{t}_i ($i=1, 2, \dots, n-1$) 构造矩阵 \hat{X} , 使得 $\hat{X} \in \tilde{X}$, 并计算 $d(\hat{X}, R_\sigma)$ 。

Step10 重复 Step4~Step9, 直到 \tilde{X} 遍历模糊平移等价标准型的全体。在所有 \hat{X} 中找出使 $d(\hat{X}, R_\sigma)$ 最小的 X^σ , 之后转 Step3。

Step11 重复 Step3~Step10, 直到 σ 遍历 S_n , 从所有的 X^σ ($\sigma \in S_n$) 中找出 $X^\#$ 与 $\sigma^\# \in S_n$, 使得:

$$d(X^\#, R_{\sigma^\#}) = \inf_{\sigma \in S_n} d(X^\sigma, R_\sigma)$$

Step12 计算 $R^\# = X_{(\sigma^\#)^{-1}}^\#$, 则 $d(R^\#, R) = \inf_{X \in X_n} d(X, R)$ 。

在 FCMBP 算法中, 高阶模糊等价标准型的平移等价类数据库没有一个高效的生成算法, 并且每一个模糊等价标准型的平移等价类要定义相应的相似参数系的等价类数据库, 非常繁琐。本文提出 2 种算法解决上述问题。

3 模糊等价标准型的平移等价类数据库生成算法

根据定理 7 和定理 9, 本文考虑任意一个 n 阶模糊等价矩阵的置换等价标准型。左下角的参数 t_1 组成的子阵只有 $[n/2]$ 种情况, 每种情况将 n 阶下三角分成 2 个低阶的等价标准型。这是由低阶向高阶自动生成库函数的算法的理论基础。

设 n 为所要生成模糊等价标准型的平移等价类数据库的阶数, $E(n)$ 表示 n 阶模糊等价标准型的平移等价类集合, 假设 n 阶以下的模糊等价标准型的平移等价类都已完全定义, 则 $E(n)$ 生成算法如下:

Step1 计算 $k=[n/2]$, 令 $i=1$ 。

Step2 取 $e_1 \in E[i]$ 。

Step3 取 $e_2 \in E[n-i]$ 。

Step4 将 e_1 中非对角线元素的角标加 1。

Step5 计算 e_1 中参数角标的最大值并将其赋予 M , 将 e_2 中非对角线元素的角标加 M 。

Step6 合并矩阵 $\begin{bmatrix} e_1 & D \\ D & e_2 \end{bmatrix}$, 将其存入 $E(n)$ 中。其中, D

是 $i \times (n-i)$ 阶元素都为 t_1 的矩阵。

Step7 返回 Step3, 重复 Setp3~Setp6, 直到 e_2 遍历了所有 $E[n-i]$ 。

Step8 返回 Step2, 重复 Setp2~Setp7, 直到 e_1 遍历了所有 $E[i]$ 。

Step9 令 $i=i+1$, 重复步骤 Setp2~Setp8, 直到 $i=k$ 。

以上算法可以进一步推广: 定义初始几个低阶的模糊等价标准型的平移等价类, 如 $E(1)=1, E(2)=\begin{bmatrix} 1 & t_1 \\ t_1 & 1 \end{bmatrix}$ 。再利用它

们做函数自身嵌套调用的终端, 解决 $E(3) \sim E(n)$ 的问题。这样就可以将模糊等价标准型的平移等价类的生成写成一个函

数, 直接得到想要的 $E(3) \sim E(n)$ 。

此函数的运算速度取决于函数初始模糊等价标准型的平移等价类定义的阶数, 阶数越高, 函数的运算速度越快。所以, 生成高阶的模糊等价标准型的平移等价类数据库可以采用分步进行的方法加快运算速度, 即先生成一系列阶数相对较低的模糊等价标准型的平移等价类, 再以这一低阶数据库为函数的初始定义来生成更高阶的数据库。

4 相似参数系等价类的生成算法

下面将根据第 3 节的数据库生成算法给出一个相似参数系等价类算法, 用以解决为每一个模糊等价标准型的平移等价类——配备相应参数系的问题。

根据定义 4 中的(4)可知, n 阶模糊等价标准型的平移等价类的参数是 $t_1 \sim t_{n-1}$ 。再根据定义 4 的(2)、(3)以及第 3 节算法中得到的模糊等价标准型的平移等价类结构(本文算法保证了右方相邻子阵的元素角标大于左方子阵元素的角标, 上方相邻子阵的元素角标大于下方子阵元素的角标), 给出如下相似参数系的平移等价类数据库生成算法:

Step1 读取一个模糊等价标准型的平移等价类, 设为 X 。

Step2 比较 X 下对角阵所有上下相邻的一对非对角线元素角标的大小, 按由小到大的顺序, 两两一组储存元素。

Step3 比较 X 下对角阵所有左右相邻的一对非对角线元素角标的大小, 按由小到大的顺序, 两两一组储存元素。

Step4 将 Step2 和 Step3 中的结果合并, 并剔除重复结果。

Step5 剔除冗余的大小关系, 输出结果。

上述算法无需在构建模糊等价标准型的平移等价类数据库时对每一个模糊等价标准型的平移等价类——定义相应的参数系, 只需要在生成平移等价类数据库之后用算法生成其对应的相似参数系数据库, 从而大大减少了生成数据库的工作量。

5 实例验证

首先定义模糊等价标准型的平移等价类数据库生成算法的初始低阶数据库, 包含 $E(1)=1, E(2)=\begin{bmatrix} 1 & t_1 \\ t_1 & 1 \end{bmatrix}$, 然后建立数据库生成函数, 生成 $E(8)$ 。任取 $E(8)$ 中的一个模糊等价标准型的平移等价类, 比如:

$$\begin{bmatrix} 1 & t_2 & t_2 & t_2 & t_1 & t_1 & t_1 & t_1 & t_1 \\ t_2 & 1 & t_3 & t_3 & t_1 & t_1 & t_1 & t_1 & t_1 \\ t_2 & t_3 & 1 & t_4 & t_1 & t_1 & t_1 & t_1 & t_1 \\ t_2 & t_3 & t_4 & 1 & t_1 & t_1 & t_1 & t_1 & t_1 \\ t_1 & t_1 & t_1 & t_1 & 1 & t_6 & t_5 & t_5 & t_5 \\ t_1 & t_1 & t_1 & t_1 & t_6 & 1 & t_5 & t_5 & t_5 \\ t_1 & t_1 & t_1 & t_1 & t_5 & t_5 & 1 & t_7 & t_7 \\ t_1 & t_1 & t_1 & t_1 & t_5 & t_5 & t_7 & 1 & t_8 \\ t_1 & t_1 & t_1 & t_1 & t_5 & t_5 & t_7 & t_8 & 1 \end{bmatrix}$$

其中, $t_1 < t_2, t_2 < t_3, t_3 < t_4, t_1 < t_5, t_5 < t_6, t_5 < t_7, t_7 < t_8$, 并利用第 3 节的算法计算其参数系。

这个模糊等价标准型的相似参数系的等价类见图 2。

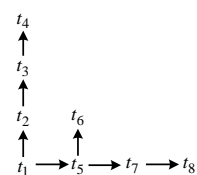


图 2 实例中相似参数系的等价类

(下转第 193 页)