第 24 卷 第 6 期
2009 年 12 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.24 No. 6
Dec. 2009

模糊聚类分析的一个改进算法及其应用

王 霞
（天津科技大学理学院，天津 300457）

摘 要：对传统模糊聚类分析方法进行研究,指出其不足之处,给出一个改进的模糊聚类分析算法,即对传递闭包进行逐行逐列改造,得到较优模糊等价矩阵,用它进行聚类比直接用传递闭包进行聚类更为合理,同时应用该方法对大学生综合素质进行评定.

关键词：聚类分析；较优等价矩阵；算法
中图分类号：O159 **文献标志码**：A **文章编号**：1672-6510(2009)06-0071-03

One Ameliorate Calculate Method and Its Use of Fuzzy Clustering Analyse

WANG Xia
(College of Science, Tianjin University of Science & Technology, Tianjin 300457, China)

Abstract: The traditional fuzzy clustering analyse method was studied. The defect of this method was pointed out and a modified algorithm was given. The transitive closure arithmetic was remade by ranks and approximate optimal fuzzy equivalent matrix arithmetic was obtained. It is better reasonable to use it than to use transitive closure arithmetic directly in clustering. The university students' comprehensive ability was evaluated by using this modified algorithm of fuzzy analyse.

Keywords: clustering analyse; optimal equivatne marix; ameliorate calculate method

聚类就是将对象聚集成若干个类别,使得同一类别对象之间有较大的相似度,不同类别两个对象相似度较小. 聚类分析是数理统计中研究“物以类聚”的一种多元分析方法,即用数学定量地确定样品的亲疏关系,从而客观地分型划类,这种分类界限是分明的,体现了非此即彼的性质. 然而,客观事物之间界限往往是不分明的,具有亦此亦彼的性质. 1969 年由 Ruspinis 提出了模糊划分概念,基于模糊划分概念的模糊聚类分析方法主要有：传递闭包法、最大树法、编网法[1]等. 模糊聚类分析反映了对象属于不同类别的不确定程度,可以更客观地反映现实世界.

1 传统模糊聚类分析方法的不足

传统模糊聚类分析方法大多解决的都是数值属性的聚类,对于一些符号属性的聚类则无法处理；传统的最大树法和编网法从图论的角度出发,虽然很直观,但是必须画图,不适合编程应用[2]；传递闭包法需要计算相似矩阵 $\boldsymbol{R}$ 的幂,其计算量随待分类对象数目的增加而呈指数规律增加. 此外,具有相同传递闭包的模糊相似矩阵并不是唯一的,为此,传递闭包法所做的模糊聚类分析的准确性就值得怀疑,这也是传递闭包法的最大不足之处.

2 改进的模糊聚类分析算法

设 $\boldsymbol{R}$ 是一个 n 阶模糊相似矩阵，$\boldsymbol{R}'$ 是由 $\boldsymbol{R}$ 生成的模糊等价矩阵,若 $\boldsymbol{R}'$ 中元素除了 1 以外,有 k 个不同的数,若所有这 k 个数在模糊相似矩阵 $\boldsymbol{R}_1$ 中只出现两次(由于有对称性),则 $\boldsymbol{R}_1$ 是一个极小模糊相似矩阵. 处于以上 $\boldsymbol{R}'$ 和 $\boldsymbol{R}_1$ 两矩阵之间的模糊相似矩阵的传递闭包都是 $\boldsymbol{R}'$[3].

引理[4] 设 $\boldsymbol{R}'=(r_{ij}')$ 是任意一个模糊等价矩阵，$s,k\in\{1,2,\cdots,n\}$； $x\in[0,1]$,那么改变 $\boldsymbol{R}'$ 的第 s 行

收稿日期：2008-11-24；**修回日期**：2009-01-23
作者简介：王 霞（1964—），女，辽宁人，教授，wxia@tust.edu.cn.

和第 s 列的元素为

$$r_{si}'' = r_{is}'' = \begin{cases} r_{ik}' \wedge x & i \neq s \\ 1 & i = s \end{cases}$$

其他元素不变，$r_{ij}'' = r_{ji}''$，则 $\boldsymbol{R}'' = (r_{ij}'')$ 仍然是模糊等价矩阵.

定理　函数 $g(x) = \sum_{i=1}^{n} (c_i \wedge x - d_i)^2$ 在 $[0,1]$ 上有最小值 g_0，其中 c_i 和 d_i 为 $[0,1]$ 上的常数.

证明　为求 $g(x)$ 的最小值，假设 c_i 在 $[0,1]$ 上的分布是 $0 \leqslant c_1 \leqslant \cdots \leqslant c_n \leqslant 1$，分别考察 $x \in [0, c_1]$，…，$x \in [c_j, c_{j+1}]$，…，$x \in [c_n, 1]$ 时，显然，当 x 处于各个小区间段时，$g(x)$ 是一元二次函数，二次项系数为正，故有唯一的最小值. 然后取以上所有 $g(x)$ 在各区间段上的最小值的最小值为 g_0 即可.

设 $\boldsymbol{B}$ 是任意一个 n 阶模糊相似矩阵，$\boldsymbol{A}$ 是由 $\boldsymbol{B}$ 生成的 n 阶模糊等价矩阵，由引理，改变 $\boldsymbol{A}$ 的第 i 列元素后，得到的仍然是模糊等价矩阵. 再由定理，可以选择 x 来继续改进第 i 列，使得与 $\boldsymbol{B}$ 的第 i 列距离：$\sum_{j=1}^{n} (a_{ji} \wedge x - b_{ji})^2$ 达到最小，这时所得的模糊等价矩阵记做 $\boldsymbol{A}^{(i)}$，即 $\boldsymbol{B}$ 的 i–较优模糊等价矩阵.

求较优模糊等价矩阵的算法：

(1) 选择 $\boldsymbol{B}$ 的传递闭包作为算法的初始矩阵，记为 $\boldsymbol{A}^{(0)} = \boldsymbol{A} = (a_{ij})$.

(2) 用 $\boldsymbol{A}$ 计算 $\boldsymbol{B}$ 的 1–较优模糊等价矩阵 $\boldsymbol{A}^{(1)}$：计算 $g_1(x) = \sum_{i=2}^{n} (a_{i1} \wedge x - b_{i1})^2$ （$x \in [0,1]$）的最小值 m_1 和最优点 x_1；计算 $g_2(x) = \sum_{i=2}^{n} (a_{i2} \wedge x - b_{i1})^2$（$x \in [0,1]$）的最小值 m_2 和最优点 x_2，…；计算 $g_n(x) = \sum_{i=2}^{n} (a_{in} \wedge x - b_{i1})^2$ （$x \in [0,1]$）的最小值 m_n 和最优点 x_n；令 $m_s = \min\{m_k \mid 1 \leqslant k \leqslant n\}$，并确定 x_s.

按下面的方法得到等价矩阵 $\boldsymbol{A}^{(1)}$：$a_{i1}^{(1)} = a_{1s}^{(1)} = a_{is} \wedge x_s (2 \leqslant i \leqslant n)$，并且，$a_{11}^{(1)} = 1$，$a_{ij}^{(1)} = a_{ji}^{(1)} = a_{ij}$ $(j \neq 1$，a_{ij} 为 $A^{(0)}$ 中的元素）. 便得到 $\boldsymbol{A}^{(1)} = (a_{ij}^{(1)})$；显然，$|\boldsymbol{A}^{(1)} - \boldsymbol{B}| \leqslant |\boldsymbol{A}^{(0)} - \boldsymbol{B}|$.

(3) 按同样的方法从 $\boldsymbol{A}^{(1)}$ 开始计算 $\boldsymbol{B}$ 的 2–较优模糊等价矩阵 $\boldsymbol{A}^{(2)}$，$\boldsymbol{B}$ 的 3–较优模矩阵 $\boldsymbol{A}^{(3)}$，…，n –较优模糊等价矩阵 $\boldsymbol{A}^{(n)}$.

当 $\boldsymbol{B}$ 的传递闭包 $\boldsymbol{A}$ 的每一列都按照上面的方法进行改进之后，就得到了所要求的较优模糊等价矩阵 $\boldsymbol{A}' = \boldsymbol{A}^{(n)}$. 显然有

$$|\boldsymbol{A}^{(n)} - \boldsymbol{B}| \leqslant \cdots \leqslant |\boldsymbol{A}^{(2)} - \boldsymbol{B}| \leqslant |\boldsymbol{A}^{(1)} - \boldsymbol{B}| \leqslant |\boldsymbol{A}^{(0)} - \boldsymbol{B}|$$

由此可见，由以上改进算法所得的较优模糊等价矩阵 $\boldsymbol{A}'$ 是根据模糊相似矩阵 $\boldsymbol{B}$ 的特性，对传递闭包 $\boldsymbol{A}$ 进行逐行逐列改造得到的，用它进行聚类会比直接用传递闭包 $\boldsymbol{A}$ 进行聚类更为合理.

3　改进算法对大学生综合素质的评定

以德、智、体、美、劳这 5 项指标表现大学生的各方面素质，并综合考虑这 5 项指标对大学生素质进行综合评定.

表 1 是 10 位学生某学年 5 项指标的得分，试对这 10 名学生的综合素质进行评定.

表 1　学生综合素质评分

Tab.1　Evaluatethe comprehensive ability of student

编号	德	智	体	美	劳	总分
1	76	95	85	83	88	424
2	68	72	84	78	66	368
3	92	85	87	83	86	433
4	87	75	93	90	86	431
5	73	74	52	67	75	341
6	79	58	84	71	68	360
7	86	87	90	98	96	457
8	76	86	65	73	79	379
9	61	74	81	84	94	394
10	72	76	86	85	83	402

3.1　构造传递闭包模糊等价矩阵

首先，对表 1 进行平移标准差变换[5]，得到表 2.

表 2　平移标准差变换

Tab.2　Translation standard difference alternate

编号	德	智	体	美	劳
1	−0.11	1.71	0.36	0.21	0.62
2	−0.01	−0.63	0.28	−0.36	−1.68
3	1.69	0.69	0.53	0.21	0.41
4	1.12	−0.33	1.03	1.00	0.41
5	−0.45	−0.43	−2.41	−1.62	−0.74
6	0.22	−2.06	0.28	−1.16	−1.47
7	1.01	0.90	0.78	1.91	1.45
8	−0.11	0.79	−1.32	−0.93	−0.32
9	−1.80	−0.43	0.03	0.32	1.24
10	−0.56	−0.22	0.45	0.43	0.09

对表 2 进行平移极差变换，得到表 3，由此完成数据标准化.

表 3　平移极差变换

Tab.3　Translation limit difference alternate

编号	德	智	体	美	劳
1	0.48	1.00	0.81	0.52	0.73
2	0.23	0.38	0.79	0.36	0.00
3	1.00	0.73	0.85	0.52	0.67
4	0.84	0.46	1.00	0.74	0.67
5	0.39	0.43	0.00	0.00	0.30
6	0.58	0.00	0.78	0.13	0.07
7	0.81	0.79	0.93	1.00	1.00
8	0.48	0.76	0.32	0.20	0.43
9	0.00	0.43	0.71	0.55	0.93
10	0.36	0.49	0.83	0.58	0.57

3.2　构造模糊相似矩阵

由夹角余弦公式，用 C 语言实现该算法，得模糊相似矩阵 $\boldsymbol{R}$ 便可构造其模糊相似矩阵如下：

$$\begin{bmatrix}
1 & 0.82 & 0.94 & 0.91 & 0.78 & 0.64 & 0.95 & 0.95 & 0.88 & 0.95 \\
 & 1 & 0.81 & 0.86 & 0.4 & 0.83 & 0.8 & 0.7 & 0.69 & 0.89 \\
 & & 1 & 0.97 & 0.8 & 0.8 & 0.96 & 0.92 & 0.78 & 0.93 \\
 & & & 1 & 0.65 & 0.84 & 0.97 & 0.83 & 0.83 & 0.97 \\
 & & & & 1 & 0.83 & 0.72 & 0.92 & 0.52 & 0.61 \\
 & & & & & 1 & 0.7 & 0.56 & 0.52 & 0.75 \\
 & & & & & & 1 & 0.88 & 0.89 & 0.97 \\
 & & & & & & & 1 & 0.89 & 0.97 \\
 & & & & & & & & 1 & 0.92 \\
 & & & & & & & & & 1
\end{bmatrix}$$

3.3　传递闭包法求模糊等价矩阵

用传递闭包法处理上面的矩阵 $\boldsymbol{R}$，用 C 语言实现该算法，即可得到其传递闭包 $\boldsymbol{R}'$：

$$\begin{bmatrix}
1 & 0.82 & 0.94 & 0.91 & 0.78 & 0.64 & 0.95 & 0.95 & 0.88 & 0.95 \\
 & 1 & 0.89 & 0.89 & 0.89 & 0.84 & 0.89 & 0.89 & 0.89 & 0.89 \\
 & & 1 & 0.97 & 0.92 & 0.84 & 0.97 & 0.95 & 0.92 & 0.97 \\
 & & & 1 & 0.92 & 0.84 & 0.97 & 0.95 & 0.92 & 0.97 \\
 & & & & 1 & 0.84 & 0.92 & 0.92 & 0.92 & 0.92 \\
 & & & & & 1 & 0.84 & 0.84 & 0.84 & 0.84 \\
 & & & & & & 1 & 0.95 & 0.92 & 0.97 \\
 & & & & & & & 1 & 0.92 & 0.95 \\
 & & & & & & & & 1 & 0.92 \\
 & & & & & & & & & 1
\end{bmatrix}$$

3.4　构造较优模糊等价矩阵

用较优模糊等价矩阵的算法，对上面的模糊等价矩阵 $\boldsymbol{R}'$ 进行改造，用 C 语言实现该算法，可得到一个更加合理的模糊等价矩阵 $\boldsymbol{R}''$：

$$\begin{bmatrix}
1 & 0.82 & 0.94 & 0.91 & 0.78 & 0.64 & 0.95 & 0.95 & 0.88 & 0.95 \\
0.82 & 1 & 0.82 & 0.82 & 0.64 & 0.7 & 0.82 & 0.82 & 0.82 & 0.82 \\
0.92 & 0.82 & 1 & 0.92 & 0.64 & 0.7 & 0.92 & 0.9 & 0.84 & 0.92 \\
0.92 & 0.82 & 0.92 & 1 & 0.64 & 0.7 & 0.92 & 0.9 & 0.84 & 0.92 \\
0.64 & 0.64 & 0.64 & 0.64 & 1 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 \\
0.7 & 0.7 & 0.7 & 0.7 & 0.64 & 1 & 0.7 & 0.7 & 0.7 & 0.7 \\
0.92 & 0.82 & 0.9 & 0.9 & 0.64 & 0.7 & 1 & 0.9 & 0.84 & 0.95 \\
0.9 & 0.82 & 0.9 & 0.9 & 0.64 & 0.7 & 0.9 & 1 & 0.84 & 0.9 \\
0.84 & 0.82 & 0.84 & 0.84 & 0.64 & 0.7 & 0.84 & 0.84 & 1 & 0.84 \\
0.92 & 0.82 & 0.92 & 0.92 & 0.64 & 0.7 & 0.95 & 0.9 & 0.84 & 1
\end{bmatrix}$$

3.5　模糊聚类分析及其结果

最后用上面的模糊等价矩阵 $\boldsymbol{R}''$ 进行聚类分析，这就完成了对 10 位学生综合素质的评定工作. 所得结果如下所示：

当阀值为 1.0 时：{1}，{2}，{3}，{4}，{5}，{6}，{7}，{8}，{9}，{10}.

当阀值为 0.95 时：{1}，{2}，{3}，{4}，{5}，{6}，{7，10}，{8}，{9}.

当阀值为 0.92 时：{1，3，4，7，10}，{2}，{5}，{6}，{8}，{9}.

当阀值为 0.9 时：{1，3，4，7，8，10}，{2}，{5}，{6}，{9}.

当阀值为 0.84 时：{1，3，4，7，8，9，10}，{2}，{5}，{6}.

当阀值为 0.82 时：{1，2，3，4，7，8，9，10}，{5}，{6}.

当阀值为 0.7 时：{1，2，3，4，6，7，8，9，10}，{5}.

当阀值为 0.64 时：{1，2，3，4，5，6，7，8，9，10}.

其动态聚类分析如图 1 所示.

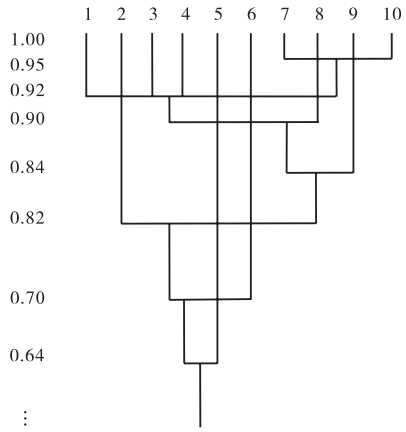

图 1　模糊聚类分析动态图

Fig.1　Dynamic table of fuzzy clustering analyse

（下转第 78 页）