

MAX- t_1 范数聚类算法的改进^{*}

闫德勤¹ 迟忠先²

(辽宁师范大学计算机系 大连116029)¹ (大连理工大学计算机系 大连116024)²

摘 要 给出了关于传递闭包模糊聚类算法与在 t_1 范数下利用模糊矩阵聚类方法的一些理论结果,同时解决了在 t_1 范数下利用模糊矩阵聚类的方法中难于确定等价类的问题,从而给出了一种基于 t_1 范数的聚类算法与应用结果。最后给出了基于 max- t_1 范数与 max- t_3 范数下利用模糊矩阵聚类方法间关系的讨论。

关键词 模糊聚类, t_1 范数, 聚类算法, 模糊矩阵

An Improvement Clustering Algorithms Based on Max- t_1 Norm

YAN De-Qin¹ CHI Zhong-Xian²

(Department of Computer Science, Liaoning Normal University, Dalian 116029)¹

(Department of Computer Science, Dalian University of Technology, Dalian 116024)²

Abstract A comparative study between transitive closure fuzzy clustering algorithm and the fuzzy clustering algorithm based on t_1 -norm with fuzzy matrix are presented. The study shows that for describing the fuzzy relation of two elements max- t_1 fuzzy transitive relation is more critical than that of max-min fuzzy transitive relation. That is the reason why t_1 -norm does not have max- t transitivity. Besides, in this paper a new algorithm to determine equivalent classes is given. With the new algorithm the fuzzy clustering algorithm based on t_1 -norm with fuzzy matrix is more reasonable. Experiments show the effectiveness of the new algorithm.

Keywords Fuzzy clustering, Max- t_1 norm, Clustering algorithm, Fuzzy matrix

1 引言

传统的传递闭包模糊聚类算法^[2~4]基于最大最小传递方式求得模糊矩阵的乘积进而求得传递闭包矩阵,选取适当的阈值对闭包矩阵截取得到一定的分类。其方法的理论基础是最小范数(又称 t_3 范数^[3,5])具有 max- t 传递性。在文[3]中 Yang 等人用事例指出了最大最小传递性的不足,并引入了 t_1 范数。同时,建立了基于模糊关系矩阵的 max- t 构成法以及求 max- t 相似矩阵和聚类的方法。虽然该方法在 t_1 范数下利用模糊矩阵聚类是对传递闭包模糊聚类算法思路的拓宽,但没有解决由于 t_1 范数不具有 max- t 传递性而对等价性的影响问题。该文讨论了传递闭包模糊聚类算法与在 t_1 范数下利用模糊矩阵聚类的方法的关系,解决了在 t_1 范数下利用模糊矩阵聚类的方法中难于确定等价类的问题,从而给出了一种基于 t_1 范数的聚类算法。最后给出了基于 max- t_1 范数与 max- t_3 范数下利用模糊矩阵聚类方法间关系的讨论。

2 递闭包聚类算法和基于 max- t 相似关系的聚类算法

考察元素的聚类性模糊理论^[1]是一种重要的工具。传递闭包聚类算法是一种基于模糊关系与模糊矩阵理论的聚类算法。矩阵的相乘使用的是最大最小关系,从范数的意义上讲是基于最小范数。Yang 等人^[3]总结了已有的相关范数,讨论了它们之间的关系,提出了一种基于 max- t 相似关系的聚类算法。以下给出相关理论以及我们的一些理论研究结果。

2.1 关于传递闭包矩阵

定义1 一个 $n \times n$ 的模糊矩阵 $A = (a_{ij})$,若有 $a_{ij} = 1 (1 \leq i \leq n)$ 则称为自反的,若 $a_{ij} = a_{ji}$ 则称为对称的。自反对称的模糊矩阵称为模糊相似矩阵。

定义2 若 $a_{ij} \geq \bigvee_k (a_{ik} \wedge a_{kj})$ 则称矩阵 $A = (a_{ij})$ 为传递的。

定义3 一个传递的相似的模糊矩阵称为模糊等价矩阵。

定义4 若对所有的 i, j 都有 $a_{ij} \leq b_{ij}$, 则称矩阵 B 包含矩阵 A , 记作 $A \subseteq B$ 。

定义5 若模糊等价矩阵 B 包含模糊相似矩阵 A , 且被所有包含矩阵 A 的模糊等价矩阵所包含, 则称 B 为 A 的等价闭包。

由定义可知相似矩阵的等价闭包即是等价闭包。一个矩阵的等价闭包即是包含该矩阵的最小模糊等价矩阵。设 R 为模糊相似矩阵, 其等价闭包矩阵由下式计算^[2,4]:

$$R^* = R \cup R^2 \cup \cdots \cup R^{n-1}.$$

又由于在矩阵相乘时使用的是最大最小导出关系, 对于模糊相似矩阵而言存在关系:

$$R \subseteq R^2 \subseteq \cdots \subseteq R^{n-1}, R^* = R^{n-1}.$$

定理1 设 A 为模糊等价矩阵, 若其 i 行(列)中最大元素在 j 列上, 如为 a_{ij} , 则 i 行(列)与 j 行(列)中除最大元素所对应的元素以及与对角元相对应的元素外, 其它相对应的元素都相同。

证明: 以行为例, 设 a_{ij} 为 i 行中最大元素, 对于由于 A 为模糊等价矩阵, 所以有 $a_{ij} \geq \min(a_{ik} \wedge a_{jk})$ 。由元素的传递性知

^{*} 辽宁师范大学校基金资助。闫德勤 博士, 教授, 主要研究领域为模式识别、图像处理等。迟忠先 教授, 博士生导师, 主要研究领域为知识发现、数据库、数据挖掘等。

这三个元素中最多只能有一个大的,其它二元必相等。若 a_{ij} 是这三个中最大的则结论成立。由于 a_{ij} 与 a_{ik} 在同一行,另外一种情况只能是 a_{jk} 最大,这意味着 $a_{ij} = a_{ik}$ 。即对应不同元素的元素为本行的最大元,从而结论成立。

2.2 关于 max-t 相似关系矩阵

以下的定义源于文[3]。

定义6(max-t 传递性) 若在 X 上的模糊关系 R 满足 $\mu_R(x, z) \geq \max_{y \in X} \{t(\mu_R(x, y), \mu_R(y, z))\}$, 则称为具有 max-t 传递性。其中, t 表示一种 t 范数。

与本文内容有关的有两种 t 范数:

$$t_1(x, y) = \max\{0, x + y - 1\}$$

$$t_3(x, y) = \min(x, y)$$

分别称为 t_1 范数和 t_3 范数。基于 max- t_3 范数的关系即是最大最小关系。

定义7(max-t 构成) 给定了一个 t 范数和初始模糊关系矩阵 $R^{(0)} = (r_{ij}^{(0)})$, 若 $R^{(n)} = (r_{ij}^{(n)})$ 依 $r_{ij}^{(n)} = \max_k \{t(r_{ik}^{(n-1)}, r_{kj}^{(n-1)})\}$, $n=1, 2, \dots$ 得出, 则此关系称为 max-t 构成。

定理2^[3] 设 R 为模糊相似矩阵, 由 max-t 构成关系所得到的矩阵有以下关系:

$$R < R^{(1)} < \dots < R^{(n)} = R^{(n+1)}$$

如果 n 不是有限的, 则 $\lim_{n \rightarrow \infty} R^{(n)} = R^{(\infty)}$, 且有:

$$R < R^{(1)} < \dots < R^{(n)} < R^{(n+1)} < \dots < R^{(\infty)}$$

定理3^[3] 设 t_1 和 t_2 为任意的两个范数, R 为模糊相似矩阵, 基于 max- t_1 构成关系有:

$$R < R^{(1)} < \dots < R^{(n)} = R^{(n+1)}$$

基于 max- t_2 构成关系有:

$$R < R^{(1)} < \dots < R^{(n)} = R^{(n+1)}$$

若 t_1 -范数 $\leq t_2$ -范数, 则 $n_1 \leq n_2$ 。

定理2中的 $R^{(n)}$ 与 $R^{(\infty)}$ 在文[3]中被称为 max-t 相似关系矩阵。

需要指出的是 max-t 相似关系矩阵的元素满足定义6的 max-t 传递性表示, 但不能依据定义1~3来说明 max-t 相似关系矩阵一定就是模糊等价矩阵。因为还要看基于 max-t 范数的模糊相似关系是否具有传递性。

定理4 若模糊相似 $n \times n$ 阶矩阵 R 基于 max- t_1 范数的相似关系矩阵为 $R^{(m)}$, 则一定有 $m \leq n-1$ 。

证明: 基于 max- t_1 范数的相似关系矩阵 $R^{(m)}$ 的得出是依据 max-t 构成关系实现的, 而 max-t 构成关系实际上相当于矩阵乘法的一种约定, 在这种约定下, $R^{(m)} = R^m$ 。若基于 max- t_3 范数, 即以 max- t_3 构成关系得出相似关系矩阵, 则这种相似关系矩阵就是相当于基于最大最小导出关系得到的传递闭包矩阵 R^s 。对于传递闭包矩阵 R^s 而言 $s \leq n-1$, 由定理3可得 $m \leq s \leq n-1$ 。

3 基于 max-t₁ 范数的聚类算法

文[3]用实例指出了在有些情况下使用 max- t_1 范数比使用 max- t_3 范数更合理。特别是对于以非距离为度量聚类问题更是如此。

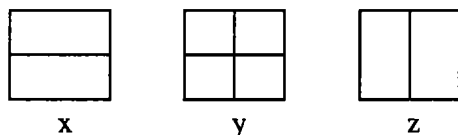


图1 x, y, z 分别表示的图形

如图1中有3个图形, 分别用 x, y, z 表示。假设 $\mu_R(x, y) = 0.7, \mu_R(y, z) = 0.7$, 则由 max-min 传递性表示得:

$$\mu_R(x, z) \geq \min\{\mu_R(x, y), \mu_R(y, z)\} = 0.7$$

由 max- t_1 传递性表示得:

$$\mu_R(x, z) \geq \max\{0, \mu_R(x, y) + \mu_R(y, z) - 1\} = 0.4$$

从 x, y, z 表示的图形可以看到, 这3个图形都有公共的方框, 其差别在于横线与竖线。从差别上看, x 与 y 差一线, 其不相相似度应为 $1 - 0.7 = 0.3$ 。同理 y 与 z 的不相似度应为 $1 - 0.7 = 0.3$ 。但 x 与 z 差两线, 显然用 max-mind 传递性表示所得到的结果不如用 max- t_1 传递性表示所得到的结果更合理。

在聚类分析中, 聚类结果的优劣在很大程度上取决于是否对聚类对象描述正确。从聚类的角度讲 max- t_1 范数是对 max- t_3 (最大最小) 范数的一个补充。因此, Yang 等人在文[3]中给出了一种基于 max- t_1 范数的聚类算法, 但该算法存在着一些问题。首先是 α 的设定标准没有依据。同时在设定 α 后若 $r_{ij} < \alpha$, 则 $r_{ij} = 0$ 的取法过于粗糙, 特别当给定聚类的分类数目时容易产生聚类错误。该算法的第二个问题是没有有效地实现出等价类的划分性。基于 max- t_3 范数的聚类方法(传递闭包聚类方法)的聚类基础是等价类的划分。而基于 max- t_1 范数的相似关系矩阵不具有等价类的划分功能, Y 算法虽然给出一种聚类方法但等价类划分的理论根据不足。该算法的第三个问题表现在文[3]中所举的一个例子上。为应用 Y 算法聚类在文[3]中给出了一个 10×10 模糊相似矩阵 $R^{(0)}$:

$$R^{(0)} = \begin{bmatrix} 1 & & & & & & & & & \\ 0.2 & 1 & & & & & & & & \\ 0.5 & 0.3 & 1 & & & & & & & \\ 0.8 & 0.6 & 0.5 & 1 & & & & & & \\ 0.6 & 0.7 & 0.3 & 0.7 & 1 & & & & & \\ 0.2 & 0.9 & 0.4 & 0.3 & 0.2 & 1 & & & & \\ 0.3 & 0.2 & 0.1 & 0.5 & 0.4 & 0.1 & 1 & & & \\ 0.9 & 0.8 & 0.3 & 0.4 & 0.5 & 0.3 & 0.6 & 1 & & \\ 0.4 & 0.3 & 0.7 & 0.1 & 0.8 & 0.7 & 0.1 & 0 & 1 & \\ 0.3 & 0.2 & 0.6 & 0.3 & 0.9 & 0.2 & 0.3 & 0.2 & 0.1 & 1 \end{bmatrix}$$

基于 max- t_1 范数的相似关系矩阵 R 为:

$$R = \begin{bmatrix} 1 & & & & & & & & & \\ 0.7 & 1 & & & & & & & & \\ 0.5 & 0.3 & 1 & & & & & & & \\ 0.8 & 0.6 & 0.5 & 1 & & & & & & \\ 0.6 & 0.7 & 0.5 & 0.7 & 1 & & & & & \\ 0.6 & 0.9 & 0.4 & 0.5 & 0.6 & 1 & & & & \\ 0.5 & 0.4 & 0.1 & 0.5 & 0.4 & 0.3 & 1 & & & \\ 0.9 & 0.8 & 0.4 & 0.7 & 0.5 & 0.7 & 0.6 & 1 & & \\ 0.4 & 0.6 & 0.7 & 0.5 & 0.8 & 0.7 & 0.2 & 0.4 & 1 & \\ 0.5 & 0.6 & 0.6 & 0.6 & 0.9 & 0.5 & 0.3 & 0.4 & 0.7 & 1 \end{bmatrix}$$

在取 $\alpha = 0.55$ 后用 Y 算法得到的聚类结果为5类: $\{1, 8, 4, 2\}, \{5, 9, 10\}, \{3\}, \{6\}, \{7\}$ 。这一结果是由于在 s3 步骤中对于 $r_{18} = r_{28} = 0.9$ 为最大值时初始聚类集合选择了 $C = \{1, 8\}$ 。若选择了 $C = \{2, 6\}$, 则最终的聚类结果为: $\{1, 2, 6, 8\}, \{5, 9, 10\}, \{3\}, \{4\}, \{7\}$ 。由于在出现多个最大值时选择初始集合具有随机性, 因此聚类的结果就不同。这样的情况出现说明该算法的鲁棒性不强。

鉴于 Y 算法存在着以上所提到的问题, 现提出一种新的基于 max- t_1 范数的聚类算法(简称 N 算法)。N 算法的核心思想就是对 max- t_1 范数的相似关系矩阵的元素做分层处理使之产生等价类。该算法不需要预先设定 α , 聚类个数可自动优化实现。

N 算法:

- s0. 设 $R^{(0)} = (r_{ij}^{(0)})$ 为 $n \times n$ 阶模糊相似矩阵, 给定初始聚类个数 c 。
 s1. 计算 $\max$ - t 相似关系矩阵 $R = (r_{ij})$ 。
 s2. 计算 $b_{ij} = 1 - r_{ij}$, 得到 $U = B = (b_{ij})_{n \times n}$ 矩阵。
 s3. 计算 $Q = \sum_{i,j=1}^n b_{ij}$, 从矩阵 B 的第 i 行找出最大元 mb , 令 $mb = \max\{mb, b_{ij}\} (i=1, 2, \dots, n)$ $d = mb / (c+1)$
 s4. for $1 \leq i \leq n$ and $1 \leq j \leq n$ do
 if $b_{ij} < d$ then $b_{ij} = 1$
 else
 $b_{ij} = 0$
 enddo
 s5. for $1 \leq i \leq n-1$ do
 if $b_{ij} = 0$ then $i = i+1$
 for $i+1 \leq j \leq n$ do
 if $b_{ij} \neq 0$ then
 第 j 行加到第 i 行
 enddo
 for $i+1 \leq j \leq n$ do
 if $b_{ij} \neq 0$ then
 第 j 列加到第 i 列
 enddo
 for $i+1 \leq j \leq n$ do
 if $b_{ij} \neq 0$ then
 把 j 归入 i 所在的一类。
 enddo
 置第 i 行与第 j 列所有元素为 0。
 enddo
 计算各聚类中心。
 s6. 若聚类数小于 c 则 $d = d/2$ 转步骤 s4, 否则合并聚类中心为 c 个。
 s8. 令 K, Z_k 分别表示聚类中心的标号和属于 k 聚类中心的元素标号。计算:

$$Q_0 = \frac{\sum_{i \in K} \sum_{j \in Z_k} b_{ij}}{\sum_{i \in K, j \in S} b_{ij}}$$

若 $Q_0 < Q$ 则取 $Q = Q_0$ 转步骤 s4。
 列出对应各聚类中心的所有元素。结束。

应用该算法对上文所列文[3]中给出的 10×10 模糊相似矩阵 $R^{(0)}$ 进行了聚类计算, 分类的结果是 $\{1, 2, 4, 6, 8\}, \{5, 9, 10\}, \{3\}, \{7\}$ 。同时, 这种分类的结果具有唯一性。

4 不同范数下的 $\max$ - t 构成

基于 $\max$ - t_3 范数的传递闭包聚类算法与基于 $\max$ - t_1 范数的聚类算法共同的基本计算结果是依据 $\max$ - t 构成产生 $\max$ - t 相似关系矩阵。其过程如下:

$$\begin{aligned} r_{ij}^{(1)} &= \max_{k_1} t(r_{ik_1}^{(0)}, r_{k_1j}^{(0)}) \\ r_{ij}^{(2)} &= \max_{k_2} t(r_{ik_2}^{(1)}, r_{k_2j}^{(1)}) = \max_{k_1, k_2} t(t(r_{ik_1}^{(0)}, r_{k_1k_2}^{(0)}), r_{k_2j}^{(0)}) \\ &\dots \end{aligned}$$

...

$$r_{ij}^{(n)} = \max_{k_1, \dots, k_n} t(\dots t(r_{ik_1}^{(0)}, r_{k_1k_2}^{(0)}), r_{k_2k_3}^{(0)}), \dots, r_{k_nj}^{(0)})$$

把相似关系矩阵 $R^{(0)} = (r_{ij}^{(0)})$ 的元素看作模糊图的模糊权值, 则上述过程就是对任意两顶间不同路长上权值的一种度量方法: $r_{ij}^{(1)}$ 是在 t 范数约束下两顶 i 与 j 间路长为 2 的最大值, $\dots, r_{ij}^{(k)}$ 是在 t 范数约束下两顶 i 与 j 间路长为 $k+1$ 的最大值。在 t_3 范数下, $\max$ - t 相似关系矩阵元素即是顶 i 到顶 j 诸路中最小权值的最大者。而在 t_1 范数下, $\max$ - t 相似关系矩阵元素即是顶 i 到顶 j 诸路权值之和减去路长个数减 1 后的最大值。显然由同一初始相似矩阵得到的 $\max$ - t_3 相似关系矩阵比相应的 $\max$ - t_1 相似关系矩阵的元素要大。以 3 个顶为例更能说明这个问题。设有 3 个顶 v_i, v_j 和 v_k , 模糊权值分别为 $r_{ik} = 0.8, r_{kj} = 0.7$ 和 $r_{ij} = 0.4$ 。由 $\max$ - t_3 构成顶 v_i 与 v_j 的权值得到提升为 0.7, 而由 $\max$ - t_1 构成顶 v_i 与 v_j 的权值则得到提升为 0.5。可见, 基于 $\max$ - t_3 范数的聚类的粗糙度比基于 $\max$ - t_1 范数的聚类的粗糙度要大。这就是为什么在文[3]中使用基于 $\max$ - t_1 构成恢复不完全数据的效果比使用 $\max$ - t_3 构成要更好的原因。

结论 文章对传递闭包聚类算法和基于 $\max$ - t_1 范数的聚类算法做了一定的理论探讨, 同时针对基于 $\max$ - t_1 范数的聚类算法的不足提出了一种新的算法。基于不同的范数实际上就是选择不同的聚类尺度和确定数据间的联系度。这方面的研究从另一角度揭示聚类的意义。

参考文献

- 1 Zadeh L A. Fuzzy sets. Inform. and Control, 1965, 8: 338~353
- 2 Tamura S, Higuchi S, Tanaka K. Pattern classification based on fuzzy relations. IEEE Trans. Systems Man Cybernet, 1978, 1: 61~66
- 3 Yang M-S, Shih H-M. Cluster analysis based on fuzzy relations. Fuzzy Sets and Systems, 2001, 120: 197~212
- 4 Lee H S. An optimal algorithm for computing the max-min transitive closure of a fuzzy similarity matrix. Fuzzy Sets and Systems, 2001, 123: 129~136
- 5 Zimmermann H J. Fuzzy set theory and its applications. Kluwer, Dordrecht, 1991

(上接第 134 页)

行的正确性。

结束语 测试预言是软件测试中一个重要部分, 一个自动化、高效的测试预言可以大大提高软件测试的效率, 缩短软件的开发进程。

本文中实现的测试预言判断程序在测试过程中是否满足时序逻辑属性, 是模型检验和软件测试的结合点。今后对该工具的改进工作包括下述几个方面:

- 1) 研究在公式分解时可以采用的优化规则, 进一步改进生成的自动机, 使得可以生成更小的、更优化的自动机。
- 2) 改进状态提取方法, 使状态的提取更为精确。
- 3) 增强对源程序的文本的分析能力, 使得能够处理更多的数据类型, 进一步支持较为复杂的程序结构。
- 4) 实现运行时监控功能。本文中采用的是宽度优先算法,

它可以每次读入一个新状态, 控制自动机向前走一步。因此, 实现运行时监控在算法上是没有问题的, 需要考虑如何把待测程序作为新的线程或进程加入整个测试程序之中, 并用适当的方法控制不同进程之间执行的同步。

参考文献

- 1 Katoen J-P. Concepts, Algorithms, and Tools for Model Checking. Friedrich-Alexander Universität Erlangen-Nürnberg, Lecture Notes of the Course "Mechanised Validation of Parallel Systems" (course number 10359) Semester 1998/1999. 47~48
- 2 Baresi L, Young M. Test Oracles. [Technical Report CIS-TR-01-02]. Aug. 2001. 6~7, 20
- 3 Dillon L D, Ramakrishna Y S. In: Proc. 4th ACM SIGSOFT Symp. Foundations of Software Engineering, San Francisco, Oct. 1996. 106~117