

基于等价相异度矩阵的聚类^{*}

赵明清^{1,2} 蒋昌俊¹ 陶树平¹

(同济大学电子与信息工程学院 上海200092)¹ (山东科技大学信息科学与工程学院 泰安271019)²

摘要 本文介绍了等价相异度矩阵的性质,证明了 $[d(i,j)]_{n \times n}^m$ 是等价相异度矩阵,并给出了等价相异度矩阵的逐次平方求解方法和基于相异度矩阵的聚类方法。最后通过实验证明了此聚类方法的可行性和有效性。

关键词 数据挖掘,等价相异度矩阵,聚类

The Clustering Based on the Equivalent Dissimilarity Degree Matrix

ZHAO Ming-Qing^{1,2} JIANG Chang-Jun¹ TAO Shu-Ping¹

(College of Electronic Information Engineering, Tongji University, Shanghai 200092)¹

(College of Information Science and Engineering, Shandong University of Science and Technology, Taian 271019)²

Abstract In this paper, the property of the dissimilarity degree matrix is discussed. It is proved that the n -th power of the n -th-order dissimilarity degree matrix is an equivalent dissimilarity degree matrix. It gives out the serial quadratic fast solution method of the equivalent dissimilarity matrix, and the clustering method based on the equivalent dissimilarity degree matrix. At last, by experiment, the clustering method is proved to be feasible and effective.

Keywords Data mining, Equivalent dissimilarity degree matrix, Clustering

1 引言

聚类是指将数据对象分组成为若干类,使得在同类中的对象间具有较高的相异度,而不同类中的对象差别较大。相异度是根据描述对象的属性值来计算的,而距离是其经常采用的度量方式。聚类分析有着广泛的应用,是数据挖掘研究领域一个非常活跃的研究课题。聚类的技术也很多,如划分方法、层次方法、基于密度的方法、基于网格的方法、基于模型的方法等^[2]。在聚类分析中,许多基于内存的聚类算法选择如下两种有代表性的数据结构:一是数据矩阵,它用 p 个属性来表示 n 个对象,表现形式是 $[x_{ij}]_{n \times p}$,其中 x_{ij} 表示对象 i 在属性 j 上的取值;二是相异度矩阵,它用来存储 n 个对象两两间的相似性,表现形式是 $[d(i,j)]_{n \times n}$,其中 $d(i,j)$ 是对象 i 和对象 j 间相异性的量化表示,其值越小,两个对象就越接近,且 $d(i,j) \geq 0, d(i,i) = 0, d(i,j) = d(j,i)$ 。从现有的聚类技术来看,各种聚类方法都没有很好地利用相异度矩阵的性质,为此,本文将定义相异度矩阵的性质,并直接从这些性质出发探讨聚类的一种新方法。

2 相异度矩阵的性质

首先,相异度矩阵是自反的(指的是 $d(i,i) = 0$)和对称的(指的是 $d(i,j) = d(j,i)$)。如果 $\forall i, j$, 都有

$$d(i,j) \leq \bigwedge_{k=1}^n (d(i,k) \vee d(k,j))$$

则称相异度矩阵 $[d(i,j)]_{n \times n}$ 是传递的。

设 $[d_1(i,j)]_{n \times n}$ 和 $[d_2(i,j)]_{n \times n}$ 是两个相异度矩阵,下面定义相异度矩阵的合成运算和并运算:

$[d_1(i,j)]_{n \times n}$ 和 $[d_2(i,j)]_{n \times n}$ 的合成是

$$[d_1(i,j)]_{n \times n} \circ [d_2(i,j)]_{n \times n} = [d_1(i,j)]_{n \times n} \vee [d_2(i,j)]_{n \times n}$$

其中

$$d(i,j) = \bigwedge_{k=1}^n (d_1(i,k) \vee d_2(k,j))$$

显然,相异度矩阵的合成运算满足交换律和结合律等。现规定:

$$[d(i,j)]_{n \times n}^1 = [d^{(1)}(i,j)]_{n \times n} = [d(i,j)]_{n \times n} [d(i,j)]_{n \times n} = [d^{(m)}(i,j)]_{n \times n} = [d(i,j)]_{n \times n}^{m-1} \circ [d(i,j)]_{n \times n} (m=2,3,\dots)$$

$[d_1(i,j)]_{n \times n}$ 和 $[d_2(i,j)]_{n \times n}$ 的并是

$$[d_1(i,j)]_{n \times n} \cup [d_2(i,j)]_{n \times n} = [d_1(i,j) \wedge d_2(i,j)]_{n \times n}$$

最后,定义相异度矩阵的包含关系:

对于 $[d_1(i,j)]_{n \times n}$ 和 $[d_2(i,j)]_{n \times n}$, 如果 $\forall i, j$, 都有 $d_1(i,j) \leq d_2(i,j)$, 则称 $[d_2(i,j)]_{n \times n}$ 包含于 $[d_1(i,j)]_{n \times n}$, 记作 $[d_2(i,j)]_{n \times n} \subseteq [d_1(i,j)]_{n \times n}$ 。

容易证明,如果 $[d_2(i,j)]_{n \times n} \subseteq [d_1(i,j)]_{n \times n}$, 那么

$$[d_2(i,j)]_{n \times n} \cup [d_1(i,j)]_{n \times n} = [d_1(i,j)]_{n \times n}$$

且对任意 $[d_3(i,j)]_{n \times n}$, 有

$$[d_2(i,j)]_{n \times n} \circ [d_3(i,j)]_{n \times n} \subseteq [d_1(i,j)]_{n \times n} \circ [d_3(i,j)]_{n \times n}$$

定理1 相异度矩阵 $[d(i,j)]_{n \times n}$ 是传递的当且仅当 $[d(i,j)]_{n \times n}^2 \subseteq [d(i,j)]_{n \times n}$ 。

证明: $\forall i, j$, 有

$$[d(i,j)]_{n \times n} \text{ 传递} \Leftrightarrow d(i,j) \leq \bigwedge_{k=1}^n (d(i,k) \vee d(k,j))$$

$$\Leftrightarrow d(i,j) \leq d^{(2)}(i,j)$$

$$\Leftrightarrow [d(i,j)]_{n \times n}^2 \subseteq [d(i,j)]_{n \times n} \quad \square$$

如果相异度矩阵 $[d(i,j)]_{n \times n}$ 满足自反性和对称性, 则称 $[d(i,j)]_{n \times n}$ 是相似的; 如果还满足传递性, 则称 $[d(i,j)]_{n \times n}$ 是等价的。自然地, $[d(i,j)]_{n \times n}$ 是相似的, 但不一定是等价的。

定理2 设 $[d(i,j)]_{n \times n}$ 是相异度矩阵, 则 $[d(i,j)]_{n \times n}^m$ 是等价的。

^{*} 基金项目: 国家自然科学基金项目(699673029, 69933020)、国家973重点基础研究发展规划项目(G1998030604)、国家杰出青年科学基金项目(60125205)。赵明清 副教授, 博士生, 研究方向为粗糙集理论与数据挖掘; 蒋昌俊 教授, 博士生导师, 研究方向为 Petri 网理论及应用、并行算法分析; 陶树平 教授, 博士生导师, 研究方向为数据仓库与数据挖掘技术。

证明:定理的证明分以下几部分:

(1)以下事实是易证的:

① $[d(i, j)]_{n \times n}^m$ 是相似的($m=1, 2, \dots$);

② $[d(i, j)]_{n \times n}^p \cup [d(i, j)]_{n \times n}^q$ 是相似的($p, q=1, 2, \dots$);

③ $\bigcup_{k=1}^{\infty} [d(i, j)]_{n \times n}^k$ 是传递的;

④ $\bigcup_{k=1}^{\infty} [d(i, j)]_{n \times n}^k$ 是等价的。

(2) $[d(i, j)]_{n \times n} \subseteq [d(i, j)]_{n \times n}^2 \subseteq \dots \subseteq [d(i, j)]_{n \times n}^{\infty} \subseteq \dots$ 因为 $\forall i, j$, 有

$$d^{(2)}(i, j) = \bigwedge_{k=1}^n (d(i, k) \vee d(k, j)) \leq d(i, j) \vee d(j, j) = d(i, j) \vee 0 = d(i, j)$$

故

$$[d(i, j)]_{n \times n} \subseteq [d(i, j)]_{n \times n}^2$$

更一般地,就是结论(2)。

(3) $[d(i, j)]_{n \times n}^{\infty} = [d(i, j)]_{n \times n}^{\infty}$

利用归纳法,可得

$$d^{(n)}(i, j) = \bigwedge_{j_1, \dots, j_{n-1}=1}^n (d(i, j_1) \vee d(j_1, j_2) \vee \dots \vee d(j_{n-1}, j))$$

因为不同的足码只有 n 个,所以 $i, j_1, \dots, j_{n-1}, j$ 必有相同的,从而可以推出

$$d^{(n)}(i, j) \geq d^{(n-1)}(i, j)$$

进而,有

$$[d(i, j)]_{n \times n} \subseteq [d^{(n-1)}(i, j)]_{n \times n}^{\infty}$$

再由(2)知,结论(3)成立。

这样一来,就有

$$\bigcup_{k=1}^{\infty} [d(i, j)]_{n \times n}^k = [d(i, j)]_{n \times n}^{\infty}$$

综上所述, $[d(i, j)]_{n \times n}^{\infty}$ 是等价相异度矩阵。□

3 聚类方法

基于等价相异度矩阵聚类的步骤:

第一步,选定相异度量方式,计算

$$d(i, j) (i=1, \dots, n; j=i+1, \dots, n);$$

第二步,利用逐次平方法,计算

$$[d(i, j)]_{n \times n} \rightarrow [d(i, j)]_{n \times n}^2 \rightarrow \dots \rightarrow [d(i, j)]_{n \times n}^{\infty}$$

直至首次出现 $2^k \geq n-1$,此时 $[d(i, j)]_{n \times n}^{\infty} = [d(i, j)]_{n \times n}^k$;

第三步,给定相异度水平 α ,如果 $d^{(n-1)}(i, j) \leq \alpha$,则对象 i 和对象 j 归为一类。

4 实验结果

考虑8个对象,3个属性的数据表(见表1,对象和属性的具体含义略去,对象用1, ..., 8表示,属性用A, B, C表示)。

表1

	A	B	C
1	1.80	0.95	0.15
2	2.10	0.99	0.21
3	3.20	1.01	0.18
4	2.20	1.03	0.17
5	2.50	0.98	0.16
6	2.80	1.02	0.20
7	1.90	1.20	0.09
8	2.00	1.30	0.11

相异度采用曼哈坦距离度量方式,计算结果:

$$[d(i, j)]_{8 \times 8} = \begin{bmatrix} 0 & 0.40 & 1.47 & 0.51 & 0.43 & 1.10 & 0.42 & 0.60 \\ & 0 & 1.14 & 0.17 & 0.46 & 0.73 & 0.53 & 0.51 \\ & & 0 & 1.02 & 0.76 & 0.43 & 0.58 & 1.56 \\ & & & 0 & 0.36 & 0.63 & 0.54 & 0.53 \\ & & & & 0 & 0.39 & 0.89 & 0.88 \\ & & & & & 0 & 1.20 & 1.17 \\ & & & & & & 0 & 0.22 \\ & & & & & & & 0 \end{bmatrix}$$

$$[d(i, j)]_{8 \times 8}^2 = \begin{bmatrix} 0 & 0.40 & 0.58 & 0.40 & 0.43 & 0.43 & 0.42 & 0.42 \\ & 0 & 0.58 & 0.17 & 0.36 & 0.46 & 0.42 & 0.51 \\ & & 0 & 0.58 & 0.43 & 0.43 & 0.58 & 0.58 \\ & & & 0 & 0.36 & 0.39 & 0.51 & 0.51 \\ & & & & 0 & 0.39 & 0.43 & 0.51 \\ & & & & & 0 & 0.58 & 0.63 \\ & & & & & & 0 & 0.22 \\ & & & & & & & 0 \end{bmatrix}$$

$$[d(i, j)]_{8 \times 8}^3 = \begin{bmatrix} 0 & 0.40 & 0.43 & 0.40 & 0.40 & 0.40 & 0.42 & 0.42 \\ & 0 & 0.43 & 0.17 & 0.36 & 0.39 & 0.42 & 0.42 \\ & & 0 & 0.43 & 0.43 & 0.43 & 0.43 & 0.51 \\ & & & 0 & 0.36 & 0.39 & 0.42 & 0.42 \\ & & & & 0 & 0.39 & 0.42 & 0.43 \\ & & & & & 0 & 0.43 & 0.43 \\ & & & & & & 0 & 0.22 \\ & & & & & & & 0 \end{bmatrix}$$

$$[d(i, j)]_{8 \times 8}^4 = \begin{bmatrix} 0 & 0.43 & 0.43 & 0.40 & 0.40 & 0.40 & 0.42 & 0.42 \\ & 0 & 0.43 & 0.17 & 0.36 & 0.39 & 0.42 & 0.42 \\ & & 0 & 0.43 & 0.43 & 0.43 & 0.43 & 0.43 \\ & & & 0 & 0.36 & 0.39 & 0.42 & 0.42 \\ & & & & 0 & 0.39 & 0.42 & 0.42 \\ & & & & & 0 & 0.42 & 0.42 \\ & & & & & & 0 & 0.22 \\ & & & & & & & 0 \end{bmatrix}$$

由此,得如图1所示的动态聚类图。

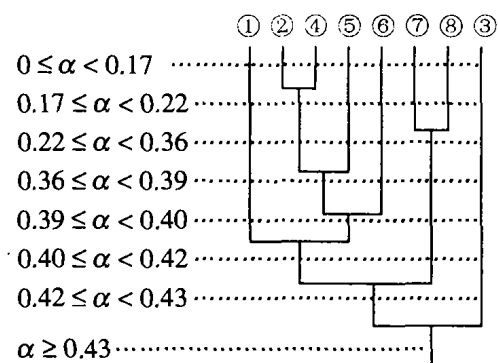


图1

结束语 实验结果表明,基于等价相异度矩阵的聚类方法是可行有效的,该方法有以下优点:①充分利用了相异度矩阵的性质;②算法容易实现;③聚类进度清楚;④对选择最好的聚类方案有指导作用;⑤聚类灵活。其缺点是在进行相异度矩阵的合成或并运算时有信息丢失,对于不典型的样本来说比较敏感。

参考文献

- 1 区奕勤,张先迪编. 模糊数学原理及应用. 成都:成都电讯工程学院出版社,1989
- 2 Han Jiawei, Kamber M 著,范明,孟小峰译. 数据挖掘—概念与技术. 北京:机械工业出版社,2001
- 3 de Sá, Marques J P 著,吴逸飞译. 模式识别—原理、方法及应用. 北京:清华大学出版社,2002
- 4 左孝凌,李为鑑编. 离散数学. 上海:上海科学技术文献出版社,1989