

高属性维稀疏数据聚类回归逻辑神经网络模型及学习算法

周永权^{1,2}, 焦李成¹

(1. 西安电子科技大学智能信息处理研究所, 陕西西安 710071; 2. 广西民族学院计算机与信息科学学院, 广西南宁 530006)

摘 要: 针对高属性维稀疏数据聚类问题, 定义了模糊取大逻辑神经元, 给出一种新的单层离散型回归逻辑神经网络模型, 由稀疏特征差异度组成的相似阵作为该网络的初始权矩阵, 通过单层离散型回归逻辑神经网络学习算法, 可求出相似矩阵的等价阵, 根据等价阵, 给定不同的阈值, 可动态地、有效地实现对高属性维稀疏数据的归并, 使得聚类结果更符合实际情况, 聚类质量较高. 相比同类聚类算法, 它具有学习、修正和应变功能, 适用于大规模稀疏数据库和稀疏数据仓库的聚类分析.

关键词: 稀疏特征; 差异度; 逻辑回归神经网络; 动态聚类法; 学习算法

中图分类号: TP311.13 **文献标识码:** A **文章编号:** 0372-2112 (2004) 08-1342-04

High Attribute Dimensional Sparse Clustering Recurrent Logical Neural Networks Model and Learning Algorithm

ZHOU Yong-quan^{1,2}, JIAO Li-cheng¹

(1. Institute of Intelligence Information Processing, Xidian University, Xi'an, Shaanxi 710071, China;

2. College of Computer and Information Science, Guangxi University for Nationalities, Nanning, Guangxi 530006, China)

Abstract: The concepts of high attribute dimensional information system and fuzzy logical neuron are firstly proposed, and a kind of new single layer discrete recurrent neural network model with n input and n output is designed. The learning of the weight matrix be done by means of finding weight matrix of equal matrix, and elements of initials weight matrix consist of sparse feature difference degree, so a sparse feature difference degree is given. It can realize dynamic state and effective classification, and the algorithm has important use in high attribute dimensional data mining and clustering analysis.

Key words: sparse feature; difference degree; logical recurrent neural network; dynamic state clustering; learning algorithm

1 引言

聚类是数据挖掘领域最为常见的技术之一, 随着对聚类技术研究的不断深入, 新的聚类算法不断出现. 但是, 每一种聚类算法^[1~6]都有其优点和特色, 同时也存在着一定的不足, 归纳起来表现在多数聚类分析算法仅适合于处理数据量比较小的情况, 例如几十个或者上百个数据对象, 而不适合于处理数据量更大的情况; 其次, 多数聚类算法在属性维数比较低的情况下可生成质量比较高的聚类结果, 却难以应用于高属性维数据聚类的情况, 仅适用于处理各个类对象分布为球形, 各个类的大小基本相同的情况, 而对数据对象分布不规则的处理能力, 显得无能为力; 再是由于人们处在信息社会, 要求数据更新非常及时, 大多数聚类算法及时处理能力都比较弱, 缺乏一定的可伸缩性. 这些问题还等待着人们进一步解决, 只有解决了这些问题, 聚类技术真正才能在数据挖掘等领域得到广泛的应用.

针对目前聚类技术中存在的问题, 文中针对一种特定类

型的高属性维稀疏数据形式, 提出一种新的高属性维稀疏数据聚类回归神经网络模型, 该模型将回归神经网络和高属性维稀疏特征的差异度组成的相似阵有机地结合起来. 即采用回归神经网络的结构和模糊逻辑的推理机制, 给出高属性维稀疏数据聚类回归神经网络学习算法, 利用该学习算法, 处理高属性维稀疏数据聚类问题, 可取得令人满意的聚类效果.

2 基本概念

2.1 高属性维稀疏信息系统

设 $S = (U, A)$ 是一高属性维信息系统, 其中 $U = \{u_1, u_2, \dots, u_n\}$ 是个体域, $A = \{a_1, a_2, \dots, a_m\}$ 是属性集, 如果每一个体都有很大一部分属性的取值为零, 那么, 我们称该信息系统 S 为高属性维稀疏信息系统. 该系统所对应的聚类问题为高属性维稀疏聚类问题.

高属性维稀疏数据的聚类是一类特殊的数据聚类问题, 在聚类应用研究中经常会遇到这样的问题: 比如一个大型超市, 销售上千种商品, 为了分析顾客群的购买行为, 需要根据

顾客购买各种产品的情况对客户进行聚类,该问题属于高属性维信息系统的范畴;若所有的产品超市都进了货,若不了解顾客购买力,导致顾客只购买了少部分商品,而对其它商品销售几乎为零,造成商品严重积压.实际上,此时信息系统 S 称为高属性维稀疏信息系统.

2.2 稀疏特征

为了有效地实现对高属性维稀疏数据进行聚类,需要根据个体属性取值的稀疏情况对问题进行分解,将一个复杂的高属性维问题转化为相对简单的多个低属性维问题.如何度量个体属性间取值的稀疏情况,为此我们给出高属性维稀疏特征的一些概念.

假设 S 为高属性维稀疏信息系统,它有 n 个个体,描述第 i 个个体的 m 个属性分别对应于区间变量值 $x_{i1}, x_{i2}, \dots, x_{im}$,那么,引入稀疏特征判断阈值 $\theta_j, j \in \{1, 2, \dots, m\}$,将区间变量转化为二态变量,表示为 $y_{i1}, y_{i2}, \dots, y_{im}$ 的公式如下:

$$y_{ij} = \begin{cases} 1, & \text{if } x_{ij} > \theta_j \\ 0, & \text{if } x_{ij} \leq \theta_j \end{cases}$$

$y_{ij}, i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m\}$ 表明了各个个体在各属性上的稀疏情况,我们称为第 i 个个体在第 j 个属性上的稀疏特征.如果 $y_{ij} = 1$ 表明第 i 个个体在第 j 个属性上是非稀疏的;如果 $y_{ij} = 0$,表明第 i 个个体在第 j 个属性上是稀疏的.

2.3 稀疏特征差异度

一般地个体间稀疏相似性可以通过计算个体间稀疏特征的差异度来描述.差异度愈大,个体愈不相似;差异度愈小,对象就愈相似,即差异度可以反映出个体间稀疏程度.假设 a 为个体 u 和个体 v 稀疏特征取值为 1 的属性个数, b 为个体 u 稀疏特征取值为 1 而个体 v 稀疏特征取值为 0 的属性个数, c 为个体 u 稀疏特征取值为 0 而个体 v 稀疏特征取值为 1 的属性个数,个体 u 与 v 之间的稀疏特征的差异度 $d(u, v)$ 可通过下式来计算^[7]:

$$d(u, v) = \frac{u \text{ 与 } v \text{ 稀疏特征取值不同的属性个数}}{u \text{ 与 } v \text{ 取值不同或同时为 1 的属性个数}} = \frac{b+c}{a+b+c} \quad (1)$$

那么,稀疏特征差异度具有以下性质:

$$0 \leq d(u, v) \leq 1; \quad d(u, u) = 0 (\text{自反性});$$

$$d(u, v) = d(v, u) \quad (\text{对称性}).$$

但不具有传递性.

2.4 相似矩阵及性质

设 S_{ij} 为个体 u_i, u_j 间的稀疏特征差异度,即 $S_{ij} = d(u_i, u_j)$,先求得两两个体间的稀疏特征差异度,然后再组成相似阵 $S = (S_{ij})_{n \times n}$,使得 S 具有以下特点:

$$S_{ii} = 0; S_{ij} = S_{ji}; i, j = 1, 2, \dots, n \quad (2)$$

称 $S = (S_{ij})_{n \times n}$ 为稀疏特征差异度组成的相似阵.

引理 相似阵 $S = (S_{ij})_{n \times n}$ 反复自乘,自乘定义为:

$$S^2 = \left(\sum_{k=1}^n (S_{ik} S_{kj}) \right)_{n \times n}; i, j = 1, 2, \dots, n$$

必定收敛,即 $S^n = S^{2^n}$, S^n 为一等阶阵^[8].

上述自乘的过程实质是将不具有传递关系的稀疏特征差异度,在 S^n 中转变成了先逻辑取小“ $\wedge$ ”,再逻辑取大“ $\vee$ ”运

算意义下具有传递性的等价关系.

3 高属性维稀疏数据聚类回归神经网络模型

1943 年,美国心理学家 McCulloch 和 Pitts 提出了神经元的数学模型——MP 模型,开创了神经网络的新时代.近年来,随着模糊系统与神经网络深入研究,人们发现原本这两个独立的领域可以相互补偿,融合在一起.于是,产生一个新的研究领域——模糊神经网络.即采用神经网络的结构和模糊逻辑的推理机制,将神经网络和模糊系统有机地结合起来.我们试图模糊回归神经网络来解决高属性维稀疏数据聚类问题,为此我们给出模糊神经网络的一些概念.

定义 设 x_i 是模糊取大逻辑神经元的 n 个输出, w_i $[0, 1]$ 是对应的权值, $i = 1, 2, \dots, n$, θ 为阈值, f 为激活函数,该模糊取大逻辑神经元输出:

$$y = f \left(\bigvee_{i=1}^n (w_i \cdot x_i) - \theta \right) = \begin{cases} 1, & \bigvee_{i=1}^n (w_i \cdot x_i) - \theta \geq 1 \\ \bigvee_{i=1}^n (w_i \cdot x_i) - \theta, & 0 \leq \bigvee_{i=1}^n (w_i \cdot x_i) - \theta \leq 1 \\ 0, & \bigvee_{i=1}^n (w_i \cdot x_i) < 0 \end{cases}$$

同理,可给出模糊逻辑取小“ $\wedge$ ”、模糊逻辑取非“ $\neg$ ”神经元的定义.由模糊逻辑神经元连接而成的网络称为模糊神经网络,它可以实现任意复杂的二值逻辑运算,因而这是逻辑神经网络的推广.

首先,考虑离散型单层回归逻辑神经网络的拓扑结构.图 1 给出 n 个逻辑神经元的单层回归神经网络,其中 $u_1, u_2, \dots, u_n$ 表示 n 个取大“ $\vee$ ”逻辑神经元; $x_1, x_2, \dots, x_n$ 既表示 n 输入变量

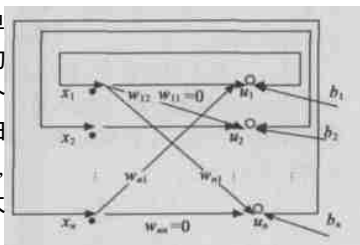


图 1 单层回归取大逻辑神经网络模型

又表示 n 个输出变量, $b_1, b_2, \dots, b_n$ 为外部输入变量,通常为网络的初值,运作后便可撤消; $w_{ij} (i, j = 1, 2, \dots, n)$ 为网络的连接权值,规定 $w_{ij} = w_{ji}$ 且 $w_{ii} = 0 (i = 1, 2, \dots, n)$; 逻辑神经元 u_i 的激活函数为 $\varphi_i (i = 1, 2, \dots, n)$,假定阈值 θ_i 已吸收到激活函数为 φ_i 之中.

对于一个时间序列 $t_0, t_1, \dots, t_k, \dots$,规定:

$$x_i(t_{k+1}) = \varphi_i \left(\bigvee_{j=1}^n (w_{ij}(t_k) \cdot \varphi_j(x_j(t_k))) \right) \quad i = 1, 2, \dots, n; k = 0, 1, 2, \dots \quad (3)$$

若把 t_k 简记为 k ,那么,上式可简记为:

$$x_i(k+1) = \varphi_i \left(\bigvee_{j=1}^n (w_{ij}(k) \cdot \varphi_j(x_j(k))) \right), \quad i = 1, 2, \dots, n; k = 0, 1, 2, \dots \quad (4)$$

其中 $w_{ij}(0) = w_{ij}^0, x_j(0) = x_j^0$ 均为初值.

特别,当激活函数 $\varphi_i (i = 1, 2, \dots, n)$ 均为恒等函数时,式(3)可转化为下列简洁式子:

$$x_i(k+1) = \bigvee_{j=1}^n (w_{ij}(k) \cdot x_j(k)), \quad k = 0, 1, 2, \dots \quad (5)$$

若记: $w(k) = (w_{ij}(k))_{n \times n}$,

$x(k) = (x_1(k), x_2(k), \dots, x_n(k))^T$,

则式(5)可写成矩阵的形式: $x(k+1) = w(k) \otimes x(k)$ (6)

其中“ $\otimes$ ”是一种关系运算符,类似于矩阵的乘法运算法则。把式(6)按 k 递返回去,便有:

$x(k+1) = w(k) \otimes w(k-1) \dots w(1) \otimes w(0) \otimes x(0)$ (7)

由式(7)可看出,此时迭代过程为权矩阵的逐步修正过程: $w(0) \rightarrow w(1) \rightarrow \dots \rightarrow w(k) \rightarrow \dots$ (8)

进一步,如果权矩阵固定初值不变(即固定的权矩阵 $w(0)$),那么式(8)又可化简为: $x(k+1) = (w(0))^{k+1} \otimes x(0)$,再把 $k+1$ 用 k 替代,便有: $x(k) = (w(0))^k \otimes x(0)$ (9)

这样一来,迭代过程转化权矩阵的乘幂过程。

4 高属性维稀疏数据聚类回归逻辑神经网络学习算法

设 S 为高属性维稀疏信息系统,它有 n 个个体,描述第 i 个个体的 m 个属性取值分别对应于区间变量值 $x_{i1}, x_{i2}, \dots, x_{im}$,稀疏判断阈值为 $\theta_j, j = \{1, 2, \dots, m\}$;描述第 i 个个体的 m 个属性的稀疏特征为 $y_{ij}, i = \{1, 2, \dots, n\}, y_{ij} \in \{0, 1\}$,那么,高属性维稀疏数据聚类方法可分四步:(1)针对给定的稀疏判断阈值为 $\theta_j, j = 1, 2, \dots, m$,计算出所有个体分别对应的各属性的稀疏特征 y_{ij} 。(2)根据稀疏特征差异度计算公式,计算出两两个体间差异度,组成相似阵 S 。(3)求相似阵 S 的等价阵,即 S 反复自乘,直到 $S^n = S^{2n}$ 为止。(4)给定任意阈值 ϵ ,对 S^n 进行等价划分。

从上面的步骤可看出,算法关键在于如何采用什么数据结构模型,快速地计算 S 的等价阵。下面可利用单层回归逻辑取大神经网络(图1)解决这一关键问题。设由稀疏特征差异度组成相似阵:

$$S = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1n} \\ S_{21} & S_{22} & \dots & S_{2n} \\ \dots & \dots & \dots & \dots \\ S_{n1} & S_{n2} & \dots & S_{nn} \end{bmatrix} \quad (10)$$

其中: $S_{ii} = 0; S_{ij} = S_{ji}; S_{ij}$ 表示第 i 个对象与第 j 对象之间的差异度,则 S 是一对称矩阵。那么结合式(5)~(8),可给出单层回归逻辑取大神经网络权值调整学习算法:

(1)置 $k := 0$; //这里 k 表示递归的步长;

(2)设 $S_{ij} := w_{ij}(0), i, j = 1, 2, \dots, n$; 为给定网络初始的权值, $S := w(0)_{n \times n}$ 为初始权矩阵;且

$x(0) = (x_1(0), x_2(0), \dots, x_n(0))^T \triangleq (b_1, b_2, \dots, b_n)^T$

为初始输入,其中 $x_j(0), j = 1, 2, \dots, n$ 分别取初始权矩阵 $w(0)_{n \times n}$ 的第 j 列, $j = 1, 2, \dots, n$; (3)根据计算公式(6)向前传播计算,得到当前时刻该网络的输出矩阵为: $x(1) = w(0) \otimes x(0)$; (4) $k := k + 1$; (5)置 $w(1) := x(1), x(1) = (x_1(1), x_2(1), \dots, x_n(1))^T$,作为下一次网络的输入,其中 $x_j(1)$ 分别取权矩阵 $w(1)_{n \times n}$ 的第 j 列, $j = 1, 2, \dots, n$; (6)重复上述递归过

程,可得到修正权值矩阵序列: $w(0) \rightarrow w(1) \rightarrow \dots \rightarrow w(k) \rightarrow \dots \rightarrow w(n) \rightarrow \dots \rightarrow w(2n) \rightarrow \dots$;

由前面的引理知,该序列一定收敛。若出现 $w(n) = w(2n)$,则学习算法结束。

通过上述步骤实施的高属性维稀疏数据聚类回归逻辑神经网络学习算法与其它聚类算法相比^[1~6],性能上至少有两点更加优越:针对大规模稀疏数据这种异常情况下的聚类问题,其它聚类算法难以实施。虽模糊等价求解方法相比经典模糊方法无实质性的改进,但它可和回归神经网络模型有机地结合起来,使用具有并行结构的回归神经网络能够使得聚类问题的解“一次学习,并行获取”,相比其它聚类算法具有快速求解、学习、修正和应变功能。

5 数值算例

假设有 8 个客户,记为 $u_i, i = 1, 2, \dots, 8$,描述每个个体的属性有 16 个,分别为该个体对 16 种产品的订购量,记为 $a_j, j = 1, 2, \dots, 16$;如下表 1 所示。以该实例为例,给出算法实施的全过程。若取稀疏判断阈值 $\theta_k = 0, k = 1, 2, \dots, 16$,计算出所有个体各属性的稀疏特征 $y_{ik}, i = 1, 2, \dots, 8; k = 1, 2, \dots, 16$;可得到原始数据表 1 中的稀疏特征(见表 2)。根据稀疏特征差异度计算公式计算出个体 u_i 与 u_j 间的差异度,组成相似阵:

$$S := w(0)_{8 \times 8} = \begin{bmatrix} 0 & 0.92 & 0.14 & 0.85 & 0.63 & 0.25 & 0 & 0.14 \\ 0.92 & 0 & 0.93 & 0.11 & 0.92 & 0.93 & 0.92 & 0.93 \\ 0.14 & 0.93 & 0 & 0.86 & 0.67 & 0.13 & 0.14 & 0 \\ 0.85 & 0.11 & 0.86 & 0 & 0.92 & 0.87 & 0.85 & 0.86 \\ 0.63 & 0.92 & 0.67 & 0.92 & 0 & 0.56 & 0.63 & 0.67 \\ 0.25 & 0.93 & 0.13 & 0.87 & 0.56 & 0 & 0.25 & 0.13 \\ 0 & 0.92 & 0.14 & 0.85 & 0.63 & 0.25 & 0 & 0.14 \\ 0.14 & 0.93 & 0 & 0.86 & 0.67 & 0.13 & 0.14 & 0 \end{bmatrix}$$

作为初始权矩阵, $x(0) = (x_1(0), x_2(0), \dots, x_8(0))^T$ 作为初始

表 1 8 个个体 16 个属性聚类问题的原始数据

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}
u_1		20				0.1	30			0.05	1			8		
u_2	20	20		20	10			20	10			20			10	
u_3		80				1	10			0.5				30		60
u_4	90	80		90	90			90	90			90		30	90	
u_5		100	100			0.1					30		100			
u_6		90				0.1	5			0.1	40		20	50		100
u_7		10				0.9	30			1	80			100		
u_8		10				0.8	100			0.9	100			90		10

表 2 8 个个体 16 个属性聚类问题的稀疏特征

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}
u_1		1				1	1			1	1			1		
u_2	1	1		1	1			1	1			1			1	
u_3		1				1	1			1				1		1
u_4	1	1		1	1			1	1			1		1	1	
u_5		1	1			1				1		1				
u_6		1				1	1			1	1		1	1		1
u_7		1				1	1			1	1			1		
u_8		1				1	1			1	1			1		1

输入,其中:

$$x_1(0) = \begin{bmatrix} 0 \\ 0.92 \\ 0.14 \\ 0.85 \\ 0.63 \\ 0.25 \\ 0 \\ 0.14 \end{bmatrix}, x_2(0) = \begin{bmatrix} 0.92 \\ 0 \\ 0.93 \\ 0.11 \\ 0.92 \\ 0.93 \\ 0.92 \\ 0.93 \end{bmatrix}, \dots, x_8(0) = \begin{bmatrix} 0.14 \\ 0.93 \\ 0 \\ 0.86 \\ 0.67 \\ 0.13 \\ 0.14 \\ 0 \end{bmatrix}$$

向前传播计算得到此时该网络输出: $x(1) = (x_1(1), x_2(1), \dots, x_8(1))^T$, 写成矩阵形式:

$$x(1) = (x_1(1), x_2(1), \dots, x_8(1))^T = \begin{bmatrix} 0.92 & 0.63 & 0.92 & 0.63 & 0.92 & 0.92 & 0.92 & 0.92 \\ 0.63 & 0.93 & 0.67 & 0.92 & 0.67 & 0.56 & 0.63 & 0.67 \\ 0.92 & 0.67 & 0.93 & 0.67 & 0.92 & 0.93 & 0.92 & 0.93 \\ 0.63 & 0.92 & 0.67 & 0.92 & 0.67 & 0.56 & 0.63 & 0.67 \\ 0.92 & 0.67 & 0.92 & 0.67 & 0.92 & 0.92 & 0.92 & 0.92 \\ 0.92 & 0.56 & 0.93 & 0.56 & 0.92 & 0.93 & 0.92 & 0.93 \\ 0.92 & 0.63 & 0.92 & 0.63 & 0.92 & 0.92 & 0.92 & 0.92 \\ 0.92 & 0.67 & 0.93 & 0.67 & 0.92 & 0.93 & 0.92 & 0.93 \end{bmatrix}$$

置 $w(1) := x(1)$, 重复上述步骤, 可得到修正后的权值矩阵:

$$w(1)_{8 \times 8} = \begin{bmatrix} 0.92 & 0.67 & 0.92 & 0.67 & 0.92 & 0.92 & 0.92 & 0.92 \\ 0.67 & 0.93 & 0.67 & 0.92 & 0.67 & 0.67 & 0.67 & 0.67 \\ 0.92 & 0.67 & 0.93 & 0.67 & 0.92 & 0.93 & 0.92 & 0.93 \\ 0.67 & 0.92 & 0.67 & 0.92 & 0.67 & 0.67 & 0.67 & 0.67 \\ 0.92 & 0.67 & 0.92 & 0.67 & 0.92 & 0.92 & 0.92 & 0.92 \\ 0.92 & 0.67 & 0.93 & 0.67 & 0.92 & 0.93 & 0.92 & 0.93 \\ 0.92 & 0.67 & 0.92 & 0.67 & 0.92 & 0.92 & 0.92 & 0.92 \\ 0.92 & 0.67 & 0.93 & 0.67 & 0.92 & 0.93 & 0.92 & 0.93 \end{bmatrix}$$

同理可得到矩阵: $w(2)_{8 \times 8}$, 且满足: $w(2)_{8 \times 8} = w(1)_{8 \times 8}$, 已收敛, $w(1)_{8 \times 8}$ 为 S 的收敛阵. 这样, 就完成了对单层回归神经网络权值的修正. 在此基础上, 若取阈值 $\theta = 0.92$, 可得到如下划分阵 (大于等于 θ 者转化为 1, 小于 θ 者转化为 0):

$$S_{0.92}^4 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

据等价划分结果, 进行归并, 得到稀疏数据的一聚类结果: $T_{0.92} = \{(u_1, u_3, u_5, u_6, u_7, u_8), (u_2, u_4)\}$; 若取阈值 $\theta = 0.93$, 可获得另一聚类结果: $T_{0.93} = \{(u_1, u_4, u_5, u_7), u_2, (u_3, u_8), u_6\}$; 同理, 若取阈值 $\theta = 0.67$, 可获得另一聚类结果: $T_{0.68} = \{(u_1, (u_2, u_4), (u_3, u_5, u_6, u_7, u_8))\}$; 可继续改变阈值, 以获得不同的聚类结果, 来适合不同问题的需要, 这是十分方便的.

6 结论

针对高属性维稀疏数据聚类问题, 提出一种新的单层离散型回归神经网络模型, 由稀疏特征差异度组成的相似阵作为网络的初始权矩阵, 通过单层离散型回归神经网络学习算法, 求出相似矩阵的等价阵, 在此基础上, 给定不同的阈值, 可动态地、有效地实现对高属性维稀疏数据的归并, 因而是使得聚类结果更符合实际情况, 聚类质量比较高, 相比目前高维数据聚类算法, 它是一种具有学习、修正和应变功能的算法, 使用具有并行结构的回归神经网络能够使得聚类问题的解“一次学习, 并行获取”, 适用于大规模稀疏数据库和稀疏数据仓库的聚类分析.

参考文献:

- [1] Zhang T, et al. BIRCH: An efficient data clustering method for very large databases[A]. Proc. of the ACM SIGMOD Int'l Conf on Management of Data[C]. Montreal: ACM press, 1996. 73 - 84.
- [2] Guha S, et al. CURE: An efficient clustering algorithm for large databases[A]. Proc. of the ACM SIGMOD Int'l Conf on Management of data[C]. Seattle: ACM Press, 1998. 73 - 84.
- [3] Guha S, et al. A robust clustering algorithm for categorical attributes[A]. Proc. of the 15th IEEE Int'l Conf on data Engineering[C]. Sydney, Australia, 1999. 512 - 521.
- [4] Ester M, et al. A density-based algorithm for discovering clusters in large spatial database with noise[A]. Proc. of 2nd Int'l Conf on KDD '96[C]. Portland: AAAI Press, 1996. 226 - 231.
- [5] 周水庚, 周傲英, 曹晶, 胡运发. 一种基于密度的快速聚类算法[J]. 计算机研究与发展, 2000, 37(11): 1287 - 1292.
- [6] Zhang W, et al. STING: A statistical information grid approach to spatial data mining[A]. Proc. of the 23th VLDB Conf[C]. Athens: Morgan Kaufmann, 1997. 186 - 195.
- [7] 武森, 高学东, M 巴斯蒂安. 高维稀疏聚类知识发现[M]. 北京: 冶金工业出版社, 2003.
- [8] (日) 浅居喜代治, 等. 模糊系统入门[M]. 赵汝怀译. 北京: 北京师范大学出版社, 1982.
- [9] 王实, 高文. 数据挖掘中的聚类方法[J]. 计算机科学, 2000, 27(4): 42 - 45.
- [10] 焦李成. 神经网络计算[M]. 西安: 西安电子科技大学出版社, 1996.

作者简介:



周永权 男, 1962 年 5 月生于陕西旬邑, 博士, 教授, 主要研究方向为神经网络, 数据挖掘, 计算智能.

焦李成 男, 1959 年 10 月生于陕西白水, 教授, 博士生导师, IEEE 高级会员, 主要研究方向为神经网络、智能信息处理、非线性科学等.