一种基于数据场的层次聚类方法

涂文燕[1], 李德毅[2], 王建民[1]
(1. 清华大学计算机系, 北京 100084; 2. 电子系统工程研究所, 北京 100039)

摘 要: 聚类分析是统计、模式识别和数据挖掘等领域中一个非常重要的研究课题, 具有广泛的应用前景. 受物理学中场论思想的启发, 提出一种基于数据场的层次聚类方法. 该方法将物质粒子间的相互作用及其场描述方法引入抽象的数域空间, 通过模拟对象在虚拟数据场中的相互作用和运动实现数据对象的自组织层次聚集. 实验显示, 该方法不依赖于用户输入参数的仔细选择, 能够发现任意大小和密度的非球形聚类, 对噪声数据不敏感, 且具有近似线性的收敛速度.

关键词: 聚类分析; 层次聚类; 数据场
中图分类号: TP18 文献标识码: A 文章编号: 0372-2112 (2006) 02-0258-05

An Hierarchical Clustering Method Based on Data Fields

GAN Wen-yan[1], LI De-yi[2], WANG Jian-min[1]
(1. Department of Computer Science & Technology, Tsinghua University, Beijing 100084, China;
2. Institute of Electronic System Engineering, Beijing 100039, China)

Abstract Clustering is a promising application area for many fields including statistics, pattern recognition, data mining, etc. The effectiveness and efficiency of existing clustering techniques, however, is somewhat limited, owing to the huge amounts data collected in databases. According the theory of fields in physics, a hierarchical clustering method based on data fields is presented. The basic idea is that the field models is introduced to describe the virtual interaction among data objects in data space and the hierarchical partitioning of the original dataset is then performed by iteratively simulating the interaction and movement of the data objects in the fields. Experimental results show that the proposed approach not only enjoys favorite clustering quality and requires no careful parameters tuning, but also has a time complexity approximately linear with respect to the size of dataset.

Key words cluster analysis; hierarchical clustering; data field

1 引言

俗话说“物以类聚, 人以群分”, 聚类是人类认知活动的一个重要内容. 所谓聚类, 就是根据事物的某些属性, 将其划分为若干类, 使得类间相似性尽量小, 类内相似性尽量大. 相关技术在统计、模式识别、机器学习和数据挖掘等诸多领域中都有广泛的应用前景[1].

在众多聚类方法中, 层次方法是一种非常重要的聚类方法, 其基本思想是递归地对数据进行合并或分裂, 将数据集划分为嵌套的类层次结构或类谱系图. 该方法的优点在于能够得到不同粒度的多层次聚类结构, 但算法的时间复杂度比较高, 至少为 $O(n^2)$; 需要用户给定一个类合并或分裂的终止条件, 某个类合并 (或分裂) 步骤一旦完成就无法撤消, 因此, 错误的合并 (或分裂) 决策会导致低质量的聚类结果[2,3]. 例如, 经典的 Single-Link、Complete-Link 算法采用链式距离衡量类间相似性, 聚类结果存在“球形偏见”或者“链式效应”; 改进的层次聚类算法 BIRCH[4]、CURE[5]、Chamelon[6]等通过集成多种聚类技术可有效提高聚类质量, 但聚类结果严重依赖于用户参数的合理设置. 因此, 积极探索新的、更有效的层次聚类算法以实现大规模数据集的有效聚集依然是聚类研究的一个开放问题.

1977年 W. E. Wright 提出一种新颖的拟重力的层次聚类算法[7], 其基本思想是空间中的每个数据对象被视为具有单位质量的质点, 通过模拟数据对象在重力作用下的相向运动将其聚集成簇. 由于重力属于一种长程作用, 即作用力的大小随距离增长衰减较慢, 相距较远的对象间仍然

收稿日期: 2003-07-24; 修回日期: 2005-09-12
基金项目: 国家自然科学基金 (No. 60375016, No. 60496323)

存在较强的相互作用, 该算法存在明显的“球形偏见”, 不能处理噪声数据, 且算法的收敛性很差, 时间复杂度至少为 O (n^2). 针对 W right算法中存在的问题, 文献[8]提出一种改进算法. 其基本思想是引入高阶重力公式缓解“球形偏见”问题, 引入空气阻力消耗系统总能量, 确保算法的最终收敛. 改进算法有效提高了 W right算法的聚类质量, 但需要引入多个用户参数, 如空气阻力系数 C_r、高阶重力公式中的距离指数 k和对象密度 D 等等, 此外, 算法性能没有得到根本改善, 时间复杂度仍然为 O (n^2).

考虑到自然界中除了重力、电磁力等长程作用以外, 还存在核力等衰减速度较快的短程作用, 根据物理学中的场论思想, 本文将物质粒子间的相互作用及其场描述方法引入抽象的数域空间, 提出一种基于数据场的层次聚类方法. 该方法将空间中的每个对象视为具有一定质量的粒子, 其周围存在一个球形对称的虚拟引力场, 位于场中的任何对象都将受到其他对象的联合作用, 从而在整个空间上确定了一个数据场. 类似物理场的矢量强度函数和标量势函数描述, 该方法引入数据场的势函数和场强函数定义, 通过模拟对象在数据场中的相互作用和运动实现数据对象的自组织层次聚集. 实验结果显示, 该方法不依赖于用户参数的仔细选择, 能够识别任意大小和密度的非球形聚类, 对噪声数据不敏感, 且具有近似线性的时间复杂度.

本文内容安排如下: 第 2节引入数据场的概念和描述方法; 第 3节给出基于数据场的层次聚类思想; 第 4节给出算法的测试比较结果; 第 5节对全文进行总结并指出今后的研究方向.

2　数据场的引入

场的概念最早是 1837年由英国物理学家法拉第提出, 用于描述物质粒子间的非接触相互作用. 随着场论思想的发展, 人们将其抽象为一个数学概念, 用来描述某个物理量或数学函数在空间内的分布规律. 如果所讨论的物理量或数学函数在空间不同点仅有数量上的区别, 则称相应的场为标量场. 反之, 如果所讨论的物理量或数学函数在空间不同点不仅有数量上的差异, 而且还有方向上的区别, 则称相应的场为矢量场. 矢量场是一种很广泛的概念, 电场、磁场、重力场和速度场等都是矢量场. 基础物理学中讨论得最多的矢量场是有源场, 一些基本的物理定律实际上都是表述有源场的分布规律, 如牛顿万有引力定律和库仑定律等. 其主要特征是将物体引入场内会受到场力的作用, 场线有起点和终点, 它们出发、终止于“源”或者“汇”. 根据场中物理量在各点处的值是否随时间变化, 有源场可以进一步分为稳定有源场和时变有源场, 稳定有源场 (又称为有势场或保守场) 具有良好的数学性质, 在一个不依赖于时间的稳定有源场中, 对应描述场的矢量强度函数 $\boldsymbol{F}(\boldsymbol{r})$, 一定存在定义在空间上的标量势函数 $\varphi(\boldsymbol{r})$, 使得两者可以通过微分算子 ∇相互联结, 即 $\boldsymbol{F}(\boldsymbol{r}) = \nabla\varphi(\boldsymbol{r})$[9].

受上述物理思想的启发, 我们尝试将物质粒子间的相互作用及其场描述方法引入抽象的数域空间. 已知空间 $\Omega \subseteq R^p$ 中包含 n 个数据对象的数据集 D, 我们将每个数据对象视为 p 维空间中具有一定质量的粒子, 其周围存在一个虚拟作用场, 位于场内的任何其他对象都将受到场力的作用, 由此所有对象的联合作用就在空间上确定了一个数据场. 根据场论知识, 对于不依赖时间的静态数据来说, 相应的数据场可以看作稳定有源场, 其空间分布规律可以采用矢量强度函数和标量势函数描述. 考虑到标量运算相对于矢量运算更简洁、直观, 我们首先引入数据场的势函数形态准则[10]:

给定空间 Ω中的数据对象 O, 令其在空间任一点 $\boldsymbol{x}$ 处产生的势值为 $\varphi(\boldsymbol{x})$, 则 $\varphi(\boldsymbol{x})$应满足:

(1) $\varphi(\boldsymbol{x})$是定义在空间 Ω上的连续、光滑、有限函数;

(2) $\varphi(\boldsymbol{x})$具有各向同性;

(3) $\varphi(\boldsymbol{x})$是对象 O 到场点 $\boldsymbol{x}$的距离的单值递减函数. 距离为 0时, $\varphi(\boldsymbol{x})$达到最大值; 距离趋于无穷大时, $\varphi(\boldsymbol{x}) \rightarrow 0$.

原则上, 符合上述准则的函数形态都可以用于定义数据场的势函数. 考虑到短程场作用更有利于揭示数据分布的聚簇特性, 本文采用具有良好数学性质的高斯函数来定义数据场的标量势:

定义 **1**　已知空间 Ω中包含 n 个对象的数据集 $D = \{x_1, x_2, \cdots, x_n\}$及其产生的数据场, 令数据对象的位置矢量分别为 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$, 则任一场点 $\boldsymbol{x}$ 处的势值和场强矢量可表示为

$$\varphi(\boldsymbol{x}) = \varphi_D(\boldsymbol{x}) = \sum_{i=1}^{n} \varphi_i(\boldsymbol{x}) = \sum_{i=1}^{n} \left(m_i \times e^{-(\|\boldsymbol{x}-\boldsymbol{x}_i\|/\sigma)^2} \right) \tag{1}$$

$$\boldsymbol{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \left((\boldsymbol{x}_i - \boldsymbol{x}) \cdot m_i \cdot e^{-(\|\boldsymbol{x}-\boldsymbol{x}_i\|/\sigma)^2} \right) \tag{2}$$

其中, $\|\boldsymbol{x} - \boldsymbol{x}_i\|$为对象 x_i 到场点 $\boldsymbol{x}$的距离; $m \geq 0 (i = 1, \cdots, n)$为对象 x_i 的质量, 满足归一化条件, 即 $\sum_{i=1}^{n} m_i = 1$; $\sigma \in (0, +\infty)$, 用于控制对象间的相互作用力程, 称为影响因子.

分析单对象数据场的场强函数及其平面力场矢量图 (如图 1所示), 可以发现, 该数据场是一个呈球形对称分

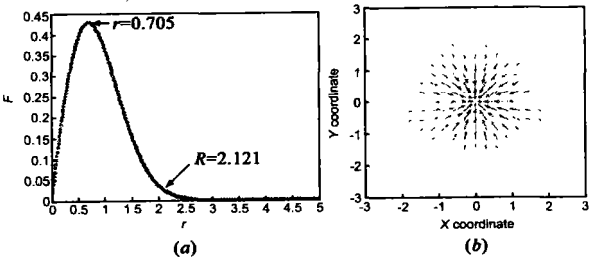

图 1　单对象数据场 (m = 1, σ = 1)

(a)场强函数; (b)场强矢量图

布的引力场, 当距场源对象的距离大于 $R \approx 2.121\sigma$ 时, 场强函数很快衰减为 0, 指示着短程场作用. 而在以场源对象位置坐标为球心、$r = 0.705\sigma$ 的球面上存在很强的向内的引力作用, 指示着数据场中单位质量的数据对象可近似看作半径为 0.705σ 的球形"质点".

3 基于数据场的层次聚类算法

已知空间 Ω 中包含 n 个对象的数据集 $D = \{x_1, x_2, \cdots, x_n\}$, 根据场论知识, 如果空间中存在多个对象且没有外力作用, 对象由于相互作用会相向运动, 并最终聚集成簇, 由此得到基于数据场的层次聚类思想. 具体实现时, 考虑到人类认知过程中初始聚类个数不是越多越好, 典型情况下, 人们至多对仅包含几十个初始聚类的类层次结构感兴趣, 因此, 算法并不显式模拟空间中所有数据对象间的相互作用和运动, 而是首先简化估计对象的质量, 得到少数质量非 0的代表对象, 然后以代表对象为聚类代表点形成数据的初始划分, 最后迭代合并代表对象实现数据的层次划分.

3.1 对象质量的简化估计

假定数据对象 $x_1, x_2, \cdots, x_n$ 取自某个 p 维连续总体, 令对象 x_i ($i = 1, \cdots, n$) 的位置矢量为 $\boldsymbol{x}_i$. 根据标量函数的叠加原理, 可以证明, 当 σ 取值一定且对象的质量满足归一化条件时, 数据场的势函数可以作为总体分布的一种估计[10]. 由此, 通过最小化势函数与总体密度间的误差平方积分可优化估计对象的质量. 具体来说, 优化目标函数可表示为一个约束条件为 $\sum_{i=1}^{n} m_i = 1$ 和 $\forall i, m_i \geq 0$ 的约束二次规划问题, 即

$$\min J = \min_{\{m_i\}} \left(\frac{1}{2 \cdot (\sqrt{2})^d} \sum_{i=1}^{n} \sum_{j=1}^{n} m_i \cdot m_j \cdot e^{-(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| / \sqrt{2}\sigma)^2} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} m_i \cdot e^{-(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| / \sigma)^2} \right) \quad (3)$$

定性分析上式的优化结果, 可以发现, 少数位于数据密集区的对象具有较大的质量, 而大多数相距较远的对象具有较小的质量或者质量为 0. 换言之, 数据场的空间分布主要取决于质量较大的对象间的相互作用, 其他大多数的对象由于质量太小对场的形成实际上不起作用. 为方便起见, 我们将估计所得的质量非 0的对象视为代表对象. 如图 2所示为数据场中 1200个数据对象的质量估计结果 ($\sigma = 0.078$), 其中 71个代表对象用删除圆圈加以标记.

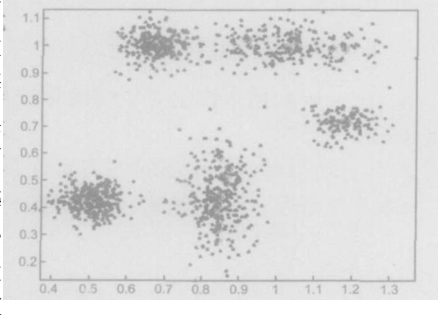

图 2 数据场中对象质量的简化估计 ($\sigma = 0.078$)

3.2 数据的初始划分

给定代表对象集 $D_{rep} \subseteq D$, 初始划分时, 每个代表对象被视为一个聚类代表点, 被同一个聚类代表点吸引的所有对象被视为一个数据类, 由此得到基于代表对象的中心聚类定义:

定义 2 已知代表对象 $x^* \in D_{rep}$ 及其质量 m^*, 如果存在子集 $C \subseteq D$, 使得 $\forall x \in C$ 都存在一个点列 $x_0 = x, x_1, \cdots, x_k \in \Omega$, 使得 x^* 与 x_k 间的距离小于等于 $0.705\sigma \cdot m^*$ 且 x_i 位于 x_{i-1} 的梯度方向 ($0 < i < k$), 则称 C 为以 x^* 为代表点的聚类.

具体实现时, 首先要剔除与所有代表对象的距离大于 3σ 的噪声数据, 然后采用场强方向指引的爬山法将其他数据对象划分给相应的代表对象, 由此得到原始数据的初始聚类划分. 初始聚类的层次合并可通过迭代模拟代表对象间的相互作用和运动来实现.

3.3 代表对象的动态合并

模拟代表对象间的相互作用和运动实现聚类合并本质上是通过迭代模拟每个代表对象在小的时间段内的运动来检查可能的类合并. 假设代表对象 $x_i^* \in D_{rep}$ 在时刻 t 的质量为 $m_i^*(t)$, 位置矢量为 $\boldsymbol{x}_i^*(t)$, $i = 1, \cdots, |D_{rep}|$, 没有外力的情况下, 对象在数据场中受到的场力和瞬间加速度可计算为

$$\boldsymbol{F}^{(t)}(x_i^*) = m_i^*(t) \cdot \sum_{x_j^* \in D_{rep}} \left(m_j^*(t) \cdot (\boldsymbol{x}_j^*(t) - \boldsymbol{x}_j^*(t)) \cdot e^{-(\|\boldsymbol{x}_j^*(t) - \boldsymbol{x}_i^*(t)\| / \sigma)^2} \right) \quad (4)$$

$$\boldsymbol{a}^{(t)}(x_i^*) = \frac{\boldsymbol{F}^{(t)}(x_i^*)}{m_i^*(t)} \quad (5)$$

如果 Δt 足够小, 每个代表对象在 $[t, t + \Delta t]$ 时间段内可近似匀变速运动. 考虑到 Wright算法的收敛性问题, 为了确保算法的最终收敛, 简单地令迭代初始时刻的速度矢量为 0, 则 $t + \Delta t$ 时刻的位置矢量可计算为

$$\boldsymbol{x}_i^*(t + \Delta t) = \boldsymbol{x}_i^*(t) + \frac{1}{2}\boldsymbol{a}^{(t)}(x_i^*) \cdot \Delta t^2 \quad (6)$$

当两个代表对象相遇或者对象间的距离小于或等于 $0.705\sigma \cdot (m_i^*(t) + m_j^*(t))$ 时, 将其合并为一个新的代表对象 x_{new}^*, 根据动量守恒定律, 新对象的质量、位置可分别计算如下

$$m_{new}^*(t) = m_i^*(t) + m_j^*(t),$$

$$\boldsymbol{x}_{new}^*(t) = \frac{m_i^*(t) \cdot \boldsymbol{x}_i^*(t) + m_j^*(t) \cdot \boldsymbol{x}_i^*(t)}{m_i^*(t) + m_j^*(t)} \quad (7)$$

算法递归执行, 直至所有的代表对象被聚合为一个对象或者满足某个用户指定的算法终止条件.

迭代过程中时间间隔 Δt 可采用自适应选取方法, 即每次迭代之前首先计算代表对象的当前最小间距 min_dist 和最大加速度 max_a, 然后取 $\Delta t = \frac{1}{f}\left(\sqrt{\frac{2\,\text{min\_dist}}{\text{max\_a}}}\right)$,

其中, 常数 f 代表时间分辨率, 通常取值 50或100.

3.4 算法描述与性能分析

基于数据场的层次聚类算法(简称为 Data-Field Clustering算法)可描述如下:

输入: 数据集 D, 影响因子 σ;

输出: 数据的层次划分 $\{\Pi_0, \Pi_1, \cdots, \Pi_k\}$;

算法步骤:

(1)M = Estimate_Mass(D, σ); //简化估计对象的质量

(2)$D_{rep} = \{x \in D \mid M(x) > 0\}$;

(3) Π_0 = Initialize_Partition(D, D_{rep}, M, σ); //初始划分

(4) $[\Pi_1, \Pi_2, \cdots, \Pi_k]$ = Dynamic_Merge(Π_0, D_{rep}, M, σ); //聚类合并

其中, 影响因子 σ 是一个重要参数, 其值的选取可采用文献[10]所给的基于熵的 σ优选方法, 即引入势熵 $H = -\sum_{i=1}^{n} \frac{\Psi_i}{Z} \log\left(\frac{\Psi_i}{Z}\right)$ 度量势场分布的不确定性, Ψ_i, i= 1, …, n为对象 x_i 的势值, $Z = \sum_{i=1}^{n} \Psi_i$ 为一个标准化因子, 通过最小势熵H 来获取最优的 σ值. 显然, 这是一个以 σ为自变量的单变量非线性函数的最优化问题, 可采用简单的一维搜索方法求解.

分析算法的时间复杂度: 步骤(1)通过求解式(3)的约束二次规划问题简化估计对象的质量, 求解约束二次规划问题存在很多标准算法, 本文采用顺序最小优化方法, 该方法的优化结果不依赖于初始点的选取, 时间复杂度为 $O(n^2)$[10]; 步骤(2)把所有质量非 0的对象划分给相应的代表对象形成初始划分, 令代表对象的个数为 n_{rep}, 平均时间复杂度为 $O(n_{rep} \cdot (n - n_{rep}))$; 步骤(3)迭代模拟代表对象在小的时间段内的运动实现聚类合并, 平均时间复杂度为 $O(n_{rep}^2)$; σ的优化涉及势熵 H 的迭代计算, 时间复杂度近似为 $O(n^2)$; 由此, 算法总的时间复杂度为 $O(n^2 + (n - n_{rep}) \cdot n_{rep} + n_{rep}^2) \approx O(n^2)$.

实际应用中, 考虑到 $O(n^2)$的时间复杂度不适用于大规模数据集的有效聚集, 可以采用样本容量 $n_{sample} \ll n$ 的随机抽样方法进一步降低参数优化和质量估计的时间开销. 具体来说, 首先从数据集 D 中随机抽取 $n_{sample} \ll n$ 个样本作为参数优化和步骤(1)的输入数据集, 然后根据代表对象集对原始数据进行初始划分, 并迭代合并代表对象形成不同层次的划分模式. 此时, 算法总的时间复杂度为 $O(n_{sample}^2 + (n - n_{rep}) \cdot n_{rep} + n_{rep}^2) \approx O(n)$. 为了确保聚类结果的有效性, 样本容量 n_{sample}的选择应尽可能保留原始数据的分布特点. 当 n 很大时, 推荐的最小抽样率[5]通常不小于 5%.

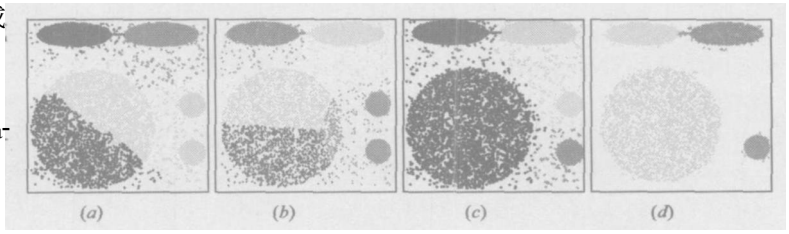

图 3　不同聚类算法的聚类结果比较. (a) BIRCH(k= 5); (b) k-means(k= 5); (c) CURE(k= 5, α= 0.3, 聚类代表点个数= 10); (d) Data-Fields Clustering (k= 5)

4 实验结果与比较

这里, 我们采用一个模拟数据集[5]来检验基于数据场的层次聚类算法的有效性. 该数据集包含 5个不同形状、大小和密度的聚类及一些均匀分布的噪声数据, 共 100,000个数据点. 考虑到 Wright算法存在收敛性问题, 而改进的重力聚类算法需要调整太多的算法参数, 我们选择流行的 K-means[1]算法和改进的层次聚类算法 BIRCH 和 CURE作为比较算法. 所有程序用 Matlab 6.1实现, 测试在一台 PC机(PIII/1G CPU、256M 内存)上进行.

测试过程中, 首先随机抽取 8000个数据点, 采用 K-means, BIRCH, CURE算法及基于数据场的层次聚类算法进行聚类分析, 聚类结果如图 3所示令聚类个数 k= 5. 显然, BIRCH 和 K-means算法存在明显的球形偏见, CURE算法在参数设置合适时能够正确发现数据分布的聚类结构, 但不能有效地处理噪声数据, 从图中可以发现, 所有的噪声数据都被划分给最邻近的聚类. 相对而言, 基于数据场的层次聚类算法不仅具有良好的聚类质量, 能够有效处理噪声数据, 而且聚类结果不依赖于用户参数的仔细选择.

利用测试数据集进一步考察基于数据场的层次聚类算法的可扩展性, 测试结果如图 4所示, 显然, 该算法具有良好的可扩展性, 算法的执行时间与数据集的大小近似线性关系.

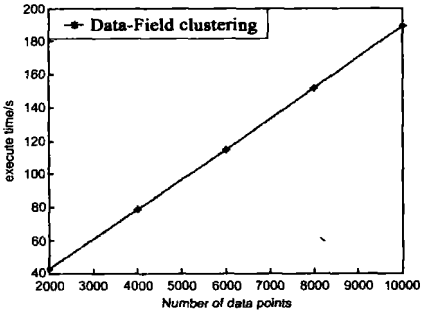

图 4　Data-Field Clustering算法的可扩展性实验

5 总结

聚类是统计、模式识别和数据挖掘等领域中一个非常基础且非常重要的研究方向, 具有极其广泛的应用前景. 受物理学中场论思想的启发, 本文提出一种基于数据场的层次聚类算法. 该方法将物质粒子间的相互作用及其场描

述方法引入抽象的数域空间,通过模拟对象在虚拟数据场中的相互作用和运动实现数据对象的自组织层次聚集.实验结果显示,该方法不依赖于用户参数的仔细选择,能够识别任意大小和密度的非球形聚类,对噪声数据不敏感,且具有近似线性的时间复杂度.目前,该算法仅在实验数据下进行了测试,今后有必要在具体应用领域中与其他聚类算法进行深入细致的比较.此外,该算法主要是针对静态数值型数据,如何对其进行扩展以便能够处理混合型数据或者动态数据,有待于进一步地研究.

参考文献:

[1] Jain A K, Murty M N, Flynn P J. Data clustering: a review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[2] Zaïane O R, Foss A, Lee C H, Wang W. On data clustering analysis: scalability, constraints and validation[A]. Proceedings of the Sixth Pacific Asia Conference on Knowledge Discovery and Data Mining[C]. Taiwan: Springer-Verlag, 2002. 28-39.

[3] Qain W N, Zhou A Y. Analyzing popular clustering algorithms from different viewpoints[J]. Journal of Software, 2002, 13(8): 1382-1394.

[4] Zhang T, Ramakrishnman R, Linvy M. BIRCH: an efficient method for very large databases[A]. Proceedings of ACM SIGMOD International Conference on Managgement of Data[C]. Canada: ACM Press, 1996. 103-114.

[5] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases[A]. Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data[C]. Seattle: ACM Press, 1998. 73-84.

[6] George K, Han E H, Kumar V. CHAMELEON: a hierarchical clustering algorithm using dynamic modeling[J]. IEEE computer, 1999, 27(3): 329-341.

[7] Wright W E. Gravitational clustering[J]. Pattern Recognition, 1977, 9(3): 151-166.

[8] Oyang Y J, Chen C Y, Yang T W. A study on the hierarchical data clustering algorithm based on gravity theory[A]. The 5th European Conference on Principles and Practive of Knowledge Discovery in Databases (PKDD 2001)[C]. Freiburg: Springer-Verlag, 2001. 350-361.

[9] Landau L D, Lifshitz E M. The classical theory of fields[M]. Beijing: Beijing World Publishing Ltd, 1999.

[10] 淦文燕.聚类-数据挖掘中的基础问题研究[D].南京:解放军理工大学, 2003.

作者简介:

淦文燕 女,1971年生于江西九江,博士后,主要研究方向为数据挖掘、数字水印,复杂网络.E-mail: wenyangan@163.com.

李德毅 男,1944年生于江苏镇江,博士生导师,中国工程院院士,主要研究方向为人工智能、数据挖掘、指挥自动化、智能控制.