文章编号:1001-506X(2010)08-1760-06

非均匀类簇密度聚类的多粒度自学习算法

曾 华[1],吴耀华[1,2],黄顺亮[3]
(1. 山东大学控制科学与工程学院,山东 济南 250061;
2. 山东大学现代物流研究中心,山东 济南 250061;
3. 山东理工大学管理学院,山东 淄博 255049)

摘 要:针对非均匀类簇密度聚类问题,从商空间粒度理论出发,提出一种多粒度自学习聚类算法(multi-granularity self-learning clustering algorithm, MSCA)。算法通过构造聚合树结构和定义粒度函数对问题逐层求解,并在每层聚合过程中根据聚合区间以自学习的方式动态确定聚合粒度,解决了传统聚类算法从非均匀类簇密度数据中无法得到不同层次的聚合特征且参数对经验依赖性过高的问题。理论和实验表明,MSCA 算法可以发现任意形状类簇,有效处理噪声,并能发现关键聚合层,具有较好的计算复杂性。

关键词:数据挖掘;聚类算法;非均匀类簇密度聚类;粒度计算;自学习算法
中图分类号:TP 181 **文献标志码**:A **DOI**:10.3969/j.issn.1001-506X.2010.08.44

Multi-granularity self-learning clustering algorithm for non-uniform cluster density

ZENG Hua[1], WU Yao-hua[1,2], HUANG Shun-liang[3]
(*1. School of Control Science and Engineering, Shandong Univ., Jinan 250061, China;*
2. The Logistics Inst., Shandong Univ., Jinan 250061, China;
3. School of Management, Shandong Univ. of Technology, Zibo 255049, China)

Abstract: Based on the quotient space granularity theory, a multi-granularity self-learning clustering algorithm (MSCA) is presented for problems with non-uniform cluster density. By constructing a feature clustering tree and defining a granularity function, MSCA solves problems layer by layer and learns clustering granularity dynamically by itself in each step. Traditional clustering algorithms with global parameters cannot discover data features in various layers, and their parameters depend on professional experience seriously, while MSCA can overcome these defects. Both theory analysis and experimental results show that MSCA can discover key clustering layers and clusters with arbitrary shape. Furthermore, it is insensitive to noise and has a satisfactory computing complexity.

Keywords: data mining; clustering algorithm; clustering with non-uniform cluster density; granular computing; self-learning algorithm

0 引 言

聚类分析是数据挖掘领域中一项重要的研究课题,它将数据集分成有意义或有用的类簇,使类簇内的数据相似度最大,类簇间的数据相似度最小。国内外研究者针对不同的数据类型和应用目的,对此进行了大量研究[1-12],并提出了许多不同的聚类算法,如基于划分的聚类算法 K-Means[2]、基于密度的聚类算法 DBSCAN[3] 和 DENCLUE[4]、基于层次的聚类算法 BIRCH[5] 和 CURE[6] 等。这些算法在类簇密度差别不大的聚类问题中具有较好的适用性。然而,现实应用的很多聚类问题中,类簇的密度具有非均匀性和不连续性,这使得上述传统聚类算法变得无能为力。另一方面,人们已不再满足于单一的聚类结果,而是越来越多地要求从数据中获取不同层面的知识,并在这些层面之间快速切换。传统聚类算法通过单一距离参数对数据进行硬性划分,这决定了它们在求解非均匀类簇密度的聚类问题时存在以下不足:全局参数直接影响聚类结果,参数的细微变化可能导致聚类结构发生巨大改变;参数的设

收稿日期:2009-04-17; **修回日期**:2009-11-10。
基金项目:国家自然科学基金(50175064)资助课题
作者简介:曾华(1981-),女,博士研究生,主要研究方向为数据挖掘与组合优化。E-mail:zenghua@mail.sdu.edu.cn

置通常依赖经验,不够客观,并且参数值的确定存在很大难度;全局参数决定了聚类结果只能反映单一层面的数据特征,无法由微观到宏观考察任意层面数据特征[1]。

商空间粒度理论利用等价类对不同粒度世界进行描述[13-14],为数据多层次分析提供了理论依据。本文从商空间粒度理论出发,针对普遍存在的非均匀类簇密度聚类问题,给出一种多粒度自学习聚类算法(multi-granularity self-learning clustering algorithm, MSCA)。MSCA 由细粒度到粗粒度对数据进行逐层聚合,每层的聚合粒度通过计算当前层次的可聚区间自学习获得,克服了上述传统聚类算法的不足,可以在非均匀类簇密度聚类问题中更全面、更客观地发现数据元素之间的结构特征,获取由微观到宏观不同层面的知识。

1　多粒度聚类和自学习聚类

1.1　多粒度聚类

定义 1　设 $\boldsymbol{D}$ 是 m 维数据的论域,称映射 $D:\boldsymbol{D}\times\boldsymbol{D}\to\boldsymbol{R}$ 是论域 $\boldsymbol{D}$ 上的距离函数,并且

$$D(x_1,x_2)=\left(\sum_{i=1}^{m}\mid x_1^{(i)}-x_2^{(i)}\mid^{p}\right)^{1/p},\ x_1,x_2\in\boldsymbol{D}\quad(1)$$

式中,$\boldsymbol{R}$ 为非负实数集,p 为距离参数。当 $p=1$ 时,D 为 Manhattan 距离;$p=2$ 时,D 为 Euclidian 距离,可根据实际需要设置。

定义 2　设 X 是论域 $\boldsymbol{D}$ 上的数据集,称集合 $\boldsymbol{X}\subseteq 2^X$ 是 X 的一个聚合域,当且仅当 $\boldsymbol{X}$ 同时满足:① 对 $\forall X_i\in\boldsymbol{X}$,有 $X_i\neq\varnothing$;② 对 $\forall X_i,X_j\in\boldsymbol{X}$ 且 $X_i\neq X_j$,有 $X_i\cap X_j=\varnothing$;③ $\bigcup_{X_i\in\boldsymbol{X}}X_i=X$。称聚合域 $\boldsymbol{X}$ 中的元素为聚合块。

定义 3　设 $\boldsymbol{X}$ 是数据集 X 的聚合域,称映射 $D':\boldsymbol{X}\times\boldsymbol{X}\to\boldsymbol{R}$ 是聚合域 $\boldsymbol{X}$ 上的距离函数,并且

$$D'(X_1,X_2)=\min_{x_1\in X_1,x_2\in X_2}(D(x_1,x_2)),\ X_1,\ X_2\in\boldsymbol{X}\quad(2)$$

其中,$\boldsymbol{R}$ 是非负实数集,D 是论域 $\boldsymbol{D}$ 上的距离函数。

定义 4　设 $\boldsymbol{X}$ 是数据集 X 的聚合域,对于给定 $\delta\geqslant 0$ 和 $X_1,X_2\in\boldsymbol{X}$,若有 $D'(X_1,X_2)\leqslant\delta$ 成立,则称 X_1 和 X_2 具有 δ-直接粒度可聚关系,其中 D' 是聚合域 $\boldsymbol{X}$ 上的距离函数。

定义 5　设 $\boldsymbol{X}$ 是数据集 X 的聚合域,对于给定 $\delta\geqslant 0$ 和 $X_1,X_2\in\boldsymbol{X}$,若存在 $\boldsymbol{X}$ 中的序列 $X'_1,X'_2,\cdots,X'_q$,其中 $X'_1=X_1$,$X'_q=X_2$,X'_i 和 X'_{i+1} 具有 δ-直接粒度可聚关系,$i=1,2,\cdots,q$,则称 X_1 和 X_2 具有 δ-粒度可聚关系。

定理 1　δ-粒度可聚关系是等价关系。

证明　设 $\boldsymbol{X}$ 是数据集 X 的聚合域,R_δ 是 $\boldsymbol{X}$ 上的一个 δ-粒度可聚关系,其中 $\delta\geqslant 0$。现证明 R_δ 具有以下性质:① 自反性:对 $\forall X_1\in\boldsymbol{X}$,由 $D'(X_1,X_1)=0\leqslant\delta$ 可知 $X_1R_\delta X_1$ 成立;② 对称性:对 $\forall X_1,X_2\in\boldsymbol{X}$,如果 $X_1R_\delta X_2$ 成立,则存在以下两种情况:(a) $D'(X_1,X_2)\leqslant\delta$,则由 $D'(X_2,X_1)=D'(X_1,X_2)\leqslant\delta$ 可知 $X_2R_\delta X_1$ 成立;(b) $D'(X_1,X_2)>\delta$,则必然存在 $\boldsymbol{X}$ 中的序列 $X_1=X'_1,X'_2,\cdots,X'_q=X_2$,使 $D'(X'_i,X'_{i+1})\leqslant\delta$ 成立($i=1,2,\cdots,q-1$),由 $D'(X'_{i+1},X'_i)=D'(X'_i,X'_{i+1})\leqslant\delta_i$ 可知 $X'_{i+1}R_\delta X'_i$ 成立,因此有 $X'_qR_\delta X'_1$ 成立,$X_2R_\delta X_1$ 得证;③ 传递性:由定义 4 和定义 5 直接可得。由此可证 R_δ 是等价关系。

定义 6　设 $\boldsymbol{X}$ 是数据集 X 的聚合域,$X_i\in\boldsymbol{X}$,R_δ 是 $\boldsymbol{X}$ 上的 δ-粒度可聚关系,称

$$[X_i]=\{X_j\mid X_j\in\boldsymbol{X},X_iR_\delta X_j\}\quad(3)$$

为 X_i 在 X 上的 δ-可聚类。

定义 7　设 $\boldsymbol{X}$ 是数据集 X 的聚合域,$X_i\in\boldsymbol{X}$,R_δ 是 $\boldsymbol{X}$ 上的 δ-粒度可聚关系,称

$$P_\delta(X_i)=\bigcup_{X_j\in[X_i]}X_j\quad(4)$$

为 X_i 在 $\boldsymbol{X}$ 上的 δ-聚合。

由定义 7 进一步可知 $P_\delta(X_i)=\{x\mid x\in X_j,X_j\in[X_i]\}$。

定理 2　聚合域中所有元素的 δ-聚合构成的集合,仍是一个聚合域。

证明　设 $\boldsymbol{X}$ 是数据集 X 的聚合域,R_δ 是 $\boldsymbol{X}$ 上的 δ-粒度可聚关系,$\boldsymbol{X}'=\{P_\delta(X_i)\mid X_i\in\boldsymbol{X}\}$,现证明 $\boldsymbol{X}'$ 满足定义 2 的 3 个条件:① 对 $\forall X'\in\boldsymbol{X}'$,$\exists X_1\in\boldsymbol{X}$ 使 $X'=P_\delta(X_1)$ 成立,由 $X_1\subseteq X'$ 且 $X_1\neq\varnothing$ 可知 $X'\neq\varnothing$;② 对 $\forall X'_1,X'_2\in\boldsymbol{X}'$ 且 $X'_1\neq X'_2$,$\exists X_1,X_2\in\boldsymbol{X}$,使 $X'_1=P_\delta(X_1)$ 且 $X'_2=P_\delta(X_2)$,下面用反证法证明 $X'_1\cap X'_2=\varnothing$:假设 $X'_1\cap X'_2\neq\varnothing$,则存在 $X_3\in\boldsymbol{X}$,使 $X_3\subseteq P_\delta(X_1)$ 且 $X_3\subseteq P_\delta(X_2)$,即 $X_1R_\delta X_3$ 且 $X_2R_\delta X_3$,由定理 1 知 $X_1R_\delta X_2$ 成立,则 $[X_1]=[X_2]$,即 $P_\delta(X_1)=P_\delta(X_2)$,于是有 $X'_1=X'_2$,这与条件 $X'_1\neq X'_2$ 相矛盾,因此必有 $X'_1\cap X'_2=\varnothing$;③ 对 $\forall x\in X$,$\exists X_1\in\boldsymbol{X}$ 使 $x\in X_1$,即 $x\in P_\delta(X_1)$ 成立,由 $P_\delta(X_1)\in\boldsymbol{X}'$ 可知 $x\in\bigcup_{X'\in\boldsymbol{X}'}X'$;反之,对 $\forall x\in\bigcup_{X'\in\boldsymbol{X}'}X'$,存在 $X'\in\boldsymbol{X}'$ 和 $X_1\in\boldsymbol{X}$ 使 $x\in X'$ 和 $X'=P_\delta(X_1)$ 成立,即存在 $X_2\in[X_1]$ 使 $x\in X_2$,由 $X_2\in\boldsymbol{X}$ 和 $X_2\subseteq X$ 可知 $x\in X$,由此可得 $\bigcup_{X'\in\boldsymbol{X}'}X'=X$。综上得证 $\boldsymbol{X}'$ 是 X 的一个聚合域。

数据集 X 的一个聚合域 $\boldsymbol{X}$ 就是 X 的一个划分,不同的 $\boldsymbol{X}$ 构成 X 不同层次上的聚类结果。

定义 8　设 $\boldsymbol{X}$ 是数据集 X 的聚合域,称三元组 $(\boldsymbol{X},\delta_{\mathrm{Min}},\delta_{\mathrm{Max}})$ 是 X 的一个粒度聚合层,并且

$$\delta_{\mathrm{Min}}=\min_{X_1,X_2\in\boldsymbol{X},X_1\neq X_2}\{D'(X_1,X_2)\}\quad(5)$$

$$\delta_{\mathrm{Max}}=\max_{X_1,X_2\in\boldsymbol{X},X_1\neq X_2}\{D'(X_1,X_2)\}\quad(6)$$

式中,D' 为 $\boldsymbol{X}$ 上的距离函数。分别称 δ_{Min} 和 δ_{Max} 是 $\boldsymbol{X}$ 上的最小可聚粒度和最大可聚粒度,称区间 $[\delta_{\mathrm{Min}},\delta_{\mathrm{Max}}]$ 是 $\boldsymbol{X}$ 的可聚区间。

定义 9　设 $(\boldsymbol{X},\delta_{\mathrm{Min}},\delta_{\mathrm{Max}})$ 是数据集 X 的一个粒度聚合层,聚合粒度 $\delta\in[\delta_{\mathrm{Min}},\delta_{\mathrm{Max}}]$,$\boldsymbol{X}'=\{P_\delta(X_i)\mid X_i\in\boldsymbol{X}\}$ 是 $\boldsymbol{X}$ 中元素的 δ-聚合构成的新聚合域,称由 $(\boldsymbol{X},\delta_{\mathrm{Min}},\delta_{\mathrm{Max}})$ 得到 $(\boldsymbol{X}',\delta'_{\mathrm{Min}},\delta'_{\mathrm{Max}})$ 的过程为在聚合粒度 δ 上的一次粒度聚合,相应地,称 $(\boldsymbol{X}',\delta'_{\mathrm{Min}},\delta'_{\mathrm{Max}})$ 是 $(\boldsymbol{X},\delta_{\mathrm{Min}},\delta_{\mathrm{Max}})$ 的高一阶粒度聚合层,$(\boldsymbol{X},\delta_{\mathrm{Min}},\delta_{\mathrm{Max}})$ 是 $(\boldsymbol{X}',\delta'_{\mathrm{Min}},\delta'_{\mathrm{Max}})$ 的低一阶粒度聚合层,其中 $P_\delta(X_i)$ 为 X_i 在 $\boldsymbol{X}$ 上的 δ-聚合。特别地,若 $\boldsymbol{X}=\{\{x_i\}\mid x_i\in X\}$,则称 $(\boldsymbol{X},\delta_{\mathrm{Min}},\delta_{\mathrm{Max}})$ 为 X 的粒度聚合底层;若 $\boldsymbol{X}=\{X\}$,

则称$(\boldsymbol{X},\delta_{\text{Min}},\delta_{\text{Max}})$为 X 的粒度聚合顶层。

每层聚合域中的聚合块对应着数据集的一个划分，也就是一个聚类结果。从粒度聚合底层到粒度聚合顶层，聚合粒度由小到大，聚合块对数据集的划分由细到粗，同时意味着学习的知识由特殊到一般，对数据特征的反映由微观到宏观。不同粒度聚合层的全体，构成数据集的多粒度聚类空间。在多粒度聚类空间中，人们可以从任意层面进行数据分析。

1.2 自学习聚类

定义 10 设 X 是论域 $\boldsymbol{D}$ 上的数据集，称映射 $f:\boldsymbol{R}\times\boldsymbol{R}\to\boldsymbol{R}$ 是 X 的粒度函数，当且仅当对 X 上任意聚合域 $\boldsymbol{X}$ 和 $\boldsymbol{X}$ 的可聚区间$[\delta_{\text{Min}},\delta_{\text{Max}}]$，有 $f(\delta_{\text{Min}},\delta_{\text{Max}})\in[\delta_{\text{Min}},\delta_{\text{Max}}]$成立，其中 $\boldsymbol{R}$ 为非负实数集。对于数据集 X 的粒度函数 f 和聚合域 $\boldsymbol{X}$，称 $\delta=f(\delta_{\text{Min}},\delta_{\text{Max}})$是粒度聚合层$(\boldsymbol{X},\delta_{\text{Min}},\delta_{\text{Max}})$的自学习聚合粒度，简称聚合粒度。

自学习聚类算法通过粒度函数获取下一层次聚合的聚合粒度，使$(\boldsymbol{X},\delta_{\text{Min}},\delta_{\text{Max}})$上的聚合粒度只与 $\boldsymbol{X}$ 对 X 的划分相关，即聚类参数完全由数据特征客观决定，摒弃了传统聚类算法依赖人为经验对数据集盲目硬性分割的做法，从而降低结果对参数的依赖性。下面介绍一种可行的粒度函数构造方法。

定义 11 设 X 是论域 $\boldsymbol{D}$ 上的数据集，f 是 X 上的一个粒度函数，对于 X 的任意粒度聚合层$(\boldsymbol{X},\delta_{\text{Min}},\delta_{\text{Max}})$，称函数

$$\Delta f(\delta_{\text{Min}},\delta_{\text{Max}})=f(\delta_{\text{Min}},\delta_{\text{Max}})-\delta_{\text{Min}} \tag{7}$$

为 X 的粒度相对增长量函数；称 $\Delta f(\delta_{\text{Min}},\delta_{\text{Max}})$的值为聚合粒度 $f(\delta_{\text{Min}},\delta_{\text{Max}})$在$(\boldsymbol{X},\delta_{\text{Min}},\delta_{\text{Max}})$上的相对增长量。称函数

$$f'(\delta_{\text{Min}},\delta_{\text{Max}})=\frac{\Delta f(\delta_{\text{Min}},\delta_{\text{Max}})}{\delta_{\text{Min}}} \tag{8}$$

为 X 的粒度相对增长率函数；称 $f'(\delta_{\text{Min}},\delta_{\text{Max}})$的值为聚合粒度 $f(\delta_{\text{Min}},\delta_{\text{Max}})$在$(\boldsymbol{X},\delta_{\text{Min}},\delta_{\text{Max}})$上的相对增长率。

$\Delta f(\delta_{\text{Min}},\delta_{\text{Max}})$的值反映了聚合粒度增长的幅度，而 $f'(\delta_{\text{Min}},\delta_{\text{Max}})$的值反映了聚合粒度增长的速度。下面考虑聚合粒度相对增长率 $f'(\delta_{\text{Min}},\delta_{\text{Max}})$取常数的情况：当 $f'(\delta_{\text{Min}},\delta_{\text{Max}})$较小时，聚合粒度相对增长缓慢，低粒度聚合层的小粒度聚类问题需要经过大量中间层次才能转化为高粒度聚合层的大粒度聚类问题；反之，当 $f'(\delta_{\text{Min}},\delta_{\text{Max}})$较大时，聚合粒度增长迅速，聚合过程进行到一定程度后将迅速到达粒度聚合顶层，导致中间必要环节遗漏。以上两种结果均是多粒度聚类分析中不愿意出现的，只希望在低粒度聚合层中聚合粒度相对增长较快，即 $f'(\delta_{\text{Min}},\delta_{\text{Max}})$较大，而在高粒度聚合层中聚合粒度相对增长较慢，即 $f'(\delta_{\text{Min}},\delta_{\text{Max}})$较小。

为此，将层次作为一项因素引入到聚合粒度相对增长率中。以自然数 $1,2,\cdots,H$ 对粒度聚合层由低到高依次标号，其中 H 是层次总数，定义如下粒度相对增长率函数

$$g'(\delta_{\text{Min}},\delta_{\text{Max}},l)=\frac{\delta_{\text{Max}}-\delta_{\text{Min}}}{\alpha l\delta_{\text{Min}}}$$
$$\alpha\geqslant 1;\ l=1,2,\cdots,H \tag{9}$$

式中，α 为复杂因子；聚合粒度相对增长率 g'和粒度聚合层序数 l 成反比，即聚合层次越低，聚合粒度相对增长越快；聚合层次越高，聚合粒度相对增长越慢。聚合粒度相对增长率 g'和复杂因子 α 成反比，即 α 越小，聚合粒度相对增长越快，得到的聚合层次越少，对问题的分析程度越简单；α 越大，聚合粒度相对增长越慢，得到的层次聚合越多，对问题的分析也就越复杂，通过调节复杂因子 α 可以控制聚类分析的粗细程度，以适应不同应用需要。进一步将式(7)和式(8)代入式(9)，可以得到引入层次因素的完整粒度函数

$$g(\delta_{\text{Min}},\delta_{\text{Max}},l)=\delta_{\text{Min}}+\frac{\delta_{\text{Max}}-\delta_{\text{Min}}}{\alpha l} \tag{10}$$

2 多粒度自学习聚类算法设计与分析

2.1 聚合节点和聚合树

聚合节点(clustering node, CN)和聚合树(clustering node tree, CN-Tree)是多粒度自学习聚类算法核心的数据组织结构。CN-Tree 是由若干 CN 构成的树型结构，每一 CN 中保存有指向其子节点的指针列表，所处层次 Layer 和三个特殊指针，这些指针分别指向其父节点，同层前驱节点和同层后继节点。特殊地，CN-Tree 叶子的子节点指针列表中只有一个指向单一数据元素的指针，即一个叶子 CN 代表一个数据元素。

CN-Tree 具有以下性质：① 树的高度 H 不小于 1；② 对于高度为 H 的 CN-Tree，其叶子层 CN 对应唯一数据元素。CN-Tree 中同层 CN 的聚合对应它们高一层次的父节点。由叶子节点层向根节点层构建 CN-Tree 的过程就是将数据集由细粒度到粗粒度逐层聚合的过程，CN-Tree 上的同层节点反映了数据集的一个划分。

2.2 MSCA 算法描述

MSCA 算法通过构造 CN-Tree，由低到高逐层检测 CN-Tree 各层节点在当前层上的 δ-粒度可聚节点，并对它们执行聚合操作，把生成的 δ-聚合作为其上一层次的节点，重复此过程，直到满足终止条件。MSCA 算法步骤如下：

步骤 1 构造空聚合树 CN_Tree，设置复杂因子 α 和算法终止类簇数 K。

步骤 2 对数据集中每一元素构造 CN_Tree 的相应叶子节点，此时每一叶子是一个类簇，层次计数器 $l=1$，当前类簇数 $k'=n$，n 为数据元素个数。

步骤 3 若满足终止条件，即当前类簇数 $k'\leqslant K$，转步骤 7；否则，继续执行步骤 4。

步骤 4 计算第 l 粒度聚合层的可聚区间$[\delta_{\text{Min}},\delta_{\text{Max}}]$，并根据式(10)中的粒度函数计算聚合粒度 δ。

步骤 5 对每一粒度聚合层 l 中任意聚合节点 CN_i：

步骤 5.1 若不存在 CN_i 的父节点，则创建 CN_i 的父节点 CN'_i；

步骤 5.2 对 CN_i 在粒度聚合层 l 中的任意 δ 直接粒度可聚节点 CN_j 执行以下操作：若存在 CN_j 的父节点

$CN_{j'}$,且 $CN_{j'} \neq CN_{i'}$,则将 $CN_{j'}$ 中所有子节点移入 $CN_{i'}$,并删除 CN_Tree 中的节点 $CN_{j'}$;否则,若不存在 CN_j 的父节点,则将 CN_j 加入 $CN_{i'}$ 的子节点列表。

步骤 6　粒度聚合层 l 计数加 1;更新当前类簇数 k';转步骤 3。

步骤 7　结束。

2.3　算法复杂性分析

2.3.1　算法时间复杂度分析

对数据规模为 n 的聚类问题,MSCA 算法步骤 1 时间复杂度为 $O(1)$;步骤 2 构造 CN_Tree 第 1 层叶节点的时间复杂度为 $O(n)$;算法运算量主要集中在步骤 3 到步骤 6 的循环中。下面分别从最坏和最好两种情况分析这一部分的时间复杂度。在最坏情况下,复杂因子 α 取值足够大,每一层次均有 $\delta \approx \delta_{\mathrm{Min}}$,且聚类数仅减 1,那么 CN_Tree 的高度为 n,即步骤 3 到步骤 6 的循环次数不超过 n。CN_Tree 第 l 层节点数为 $n-l+1$,步骤 5.2 中粒度可聚关系判断的比较次数为 $\frac{(n-l+1)(n-l)}{2}$,n 层粒度可聚关系判断的比较总次数为 $\sum_{l=1}^{n} \frac{(n-l+1)(n-l)}{2} = \frac{n(n+1)(2n+1)}{6}$;CN_Tree第 l 层步骤 5 循环中共修改指针 l 次,n 层聚合修改指针总次数为 $\sum_{l=1}^{n} l = \frac{n(n+1)}{2}$;步骤 6 赋值操作时间复杂度为 $O(1)$;对步骤 4 的可聚区间 $[\delta_{\mathrm{Min}}, \delta_{\mathrm{Max}}]$ 计算,在 $l=1$ 时复杂度为 $O(n^2)$,除此情况外,该操作均可在步骤 5 中同步完成,不增加算法时间复杂度。综上可知,最坏情况下 MSCA 的时间复杂度为 $O(n^3)$。最好情况下,每一次聚合时 CN_Tree新层次的节点数都是前一层次节点数的 $1/m$,此时 CN_Tree 为一棵高度为 $H=1+\log_m n$ 的完全 m 叉树,CN_Tree 第 l 层节点数为 m^{n-l},那么步骤 5.2 中粒度可聚关系判断中比较次数为 $\frac{m^{H-l}(m^{H-l}+1)}{2}$,$H$ 层粒度可聚关系判断的比较总次数为 $\sum_{l=1}^{H} \frac{m^{H-l}(m^{H-l}+1)}{2}$;CN_Tree 第 l 层步骤 5 修改指针次数为 m^{H-l},则 H 层聚合修改指针总次数为 $\sum_{l=1}^{H} m^{H-l} = \frac{m^H-1}{m-1} = \frac{mn-1}{m-1}$;其余操作时间复杂度与最坏情况相同。因此,对常数 $m \ll n$ 的情况,算法的时间复杂度为 $O(n^2)$。借助空间存取技术如 R*-trees[15],CN_Tree每层 n 次粒度可聚节点的空间查询可在 $O(n\log n)$时间内完成,则 H 层粒度可聚节点执行空间查询操作的总时间复杂度为 $O(n\log^2 n)$,此时 MSCA 时间复杂度为 $O(n\log^2 n)$。

2.3.2　算法空间复杂度分析

对数据规模为 n 的聚类问题,最坏情况下 CN_Tree 节点数为 $\sum_{l=1}^{n} l = \frac{n(n+1)}{2}$,每层可聚合的 δ 粒度可聚节点数不超过 1,此时 MSCA 空间复杂度为 $O(n^2)$。最好情况下 CN_Tree 节点数为 $\sum_{l=1}^{H} m^{H-l} = \frac{m^H-1}{m-1} = \frac{mn-1}{m-1}$,每层可聚合的 δ 粒度可聚节点数不超过 m。因此,对常数 $m \ll n$ 的情况,MSCA 空间复杂度为 $O(n)$。

3　实验分析

本节对 MSCA 算法性能进行测试,实验在 Intel 2.0 GHz/2.0 GB配置环境中完成,程序代码用 Visual C++(6.0)实现。实验所用数据如图 1 所示,其中 Dataset1,Dataset2和 Dataset3 为模仿文献[3]构造的常用模拟数据集,Dataset4是合成数据集。实验数据的数据特征如表 1 所示。

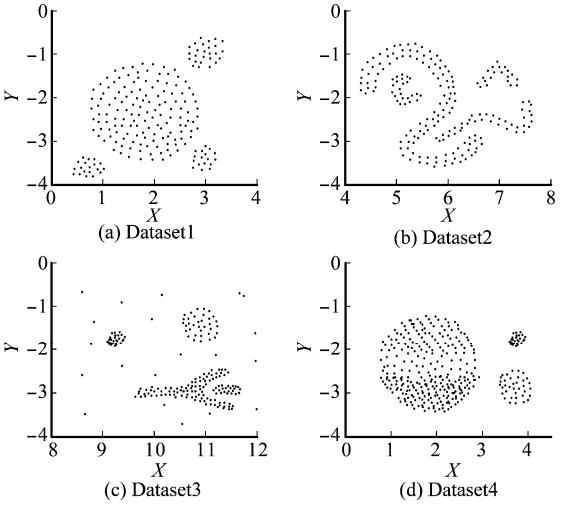

图 1　实验数据集

表 1　实验数据的数据特征

数据集	噪声比例/(%)	类簇形状	类簇密度	低密度隔离区域
Dataset1	0	球形	均匀	无
Dataset2	0	非球形	均匀	无
Dataset3	10	非球形	非均匀	无
Dataset4	0	球形	非均匀	有

3.1　算法有效性测试

用 MSCA 算法对图 1 中四组实验数据集进行多粒度聚合,参数设置为 $\alpha=10$,$k=1$。

实验得到各组数据类簇数目和聚合粒度的变化关系如图 2 所示,其中(a)、(b)和(d)三组数据由于无噪声影响,聚合粒度曲线在临近聚合结束时均有较陡的上升;而(c)因受噪声影响,聚合结束前的一个阶段内其聚合粒度上升迟缓。

为了把握聚合过程中聚类结构发生重大变化的层次,定义以下概念。

设 $\boldsymbol{X}_l$ 是第 l 聚合层上的聚合域,δ_l 是 l 层的聚合粒度,则

定义 12　l 的最大类簇扩展量为

$$I(l) = \max\left(\min_{C_i, C_j \in \boldsymbol{X}_l, C_i R_{\delta_l} C_j} (|C_i|, |C_j|)\right) \quad (11)$$

式中,C_i 为 $\boldsymbol{X}_l$ 中的类簇,$|C_i|$ 为类簇 C_i 中数据元素的个数。

定义 13　聚合层 l 的最大类簇扩展增量为

$$\Delta I(l)=\begin{cases} I(l)-I(l-1), & l>1 \\ 0, & l=1 \end{cases} \tag{12}$$

定义 14 聚合层 l 的单位粒度最大类簇扩展增量为

$$den(l)=\begin{cases} \dfrac{\Delta I(l)}{\delta_l}, & l>1 \\ 0, & l=1 \end{cases} \tag{13}$$

把 $den(l)$ 出现明显峰值的聚合层称为关键聚合层。由图 2 进一步得到图 3 的 $den(l)$ 变化关系，$den(l)$ 的峰值对应着聚合过程中的关键聚合层，这些关键聚合层反映了数据内部隐藏的聚类特征。

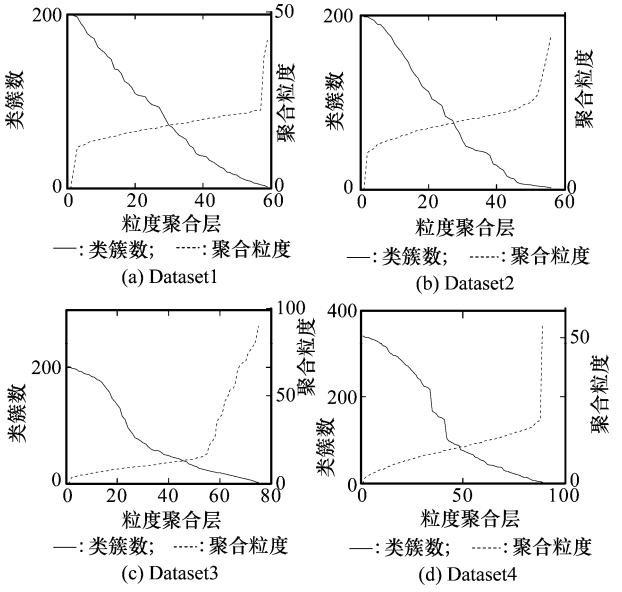

图 2 不同粒度聚合层上的类簇数和聚合粒度

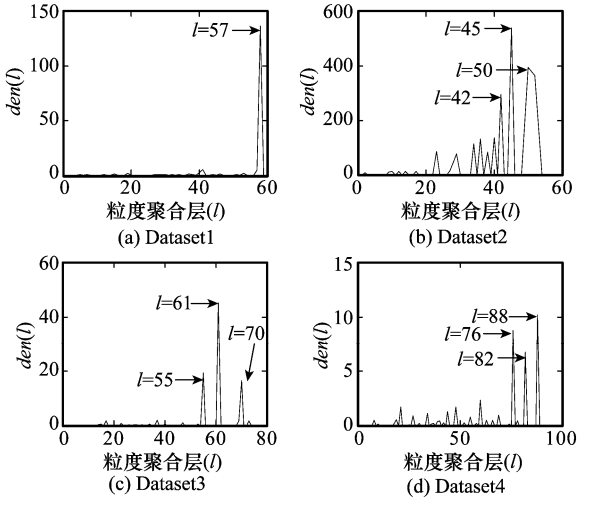

图 3 $den(l)$ 和关键聚合层

表 2 给出图 3 中关键聚合层的类簇变化和相应的聚合粒度。在这些关键聚类层上，聚合前后的聚类结果有质的区别，关键聚合层及其上的聚合粒度对描述数据特征至关重要，是多粒度聚类分析中考察的重点。

表 2 关键聚合层和相应聚合粒度

数据集	关键聚合层	聚合前类簇	聚合后类簇	聚合粒度
Dataset1	57	{C1}{C2}{C3}{C4}	{C1,C2,C3,C4}	22.2
Dataset2	42	{C1}{C2}{C3}{C4}	{C1}{C2,C3}{C4}	17.4
	45	{C1}{C2,C3}{C4}	{C1,C2,C3}{C4}	18.0
	50	{C1,C2,C3}{C4}	{C1,C2,C3,C4}	19.7
Dataset3	55	{C1}{C2}{C3}{C4}	{C1}{C2}{C3,C4}	16.7
	61	{C1}{C2}{C3,C4}	{C1}{C2,C3,C4}	38.9
	70	{C1}{C2,C3,C4}	{C1,C2,C3,C4}	69.0
Dataset4	76	{C1}{C3}{C4}{C5}	{C1,C2,C3}{C4}{C5}	17.3
	82	{C1,C2,C3}{C4}{C5}	{C1,C2,C3,C5}{C4}	18.4
	88	{C1,C2,C3,C5}{C4}	{C1,C2,C3,C4,C5}	55.0

图 4 给出图 1 中四组数据在特定关键聚合层的聚类结果，实验结果表明 MSCA 在聚类计算过程中可以准确把握类簇聚合过程，并能够根据需要发现由微观到宏观不同层次上的类簇。

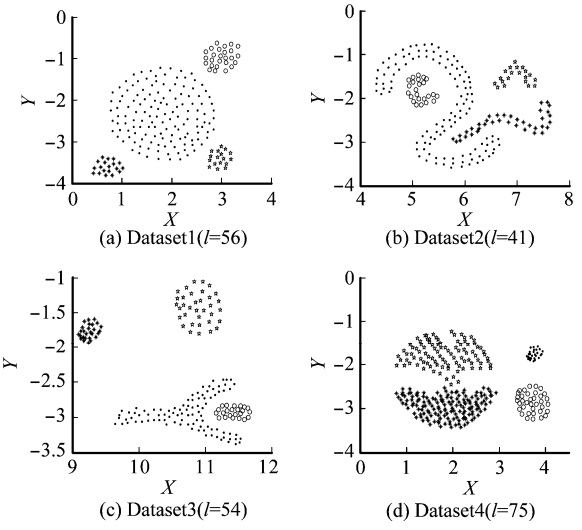

图 4 特定聚合层次上的聚类结果

3.2 与传统聚类算法的对比分析

使用图 1 中的数据集对 K-Means，DBSCAN，BIRCH，DENCLUE 和 CURE 聚类算法同文中提出的 MSCA 进行对比分析。实验中 DBSCAN、DENCLUE、BIRCH 和 CURE 算法仅当粒度参数分别接近①δ=22.2，②δ=17.4，③δ=16.7，④δ=17.3 时可以得到正确的聚类结果，故经多次实验才得以确定参数范围；而 K-Means 算法仅能较好识别具有球形聚类特征的图 1(a)，无法区分图 1 其余三组非球形聚类；MSCA 根据数据特点以自学习的方式确定聚合粒度，很好地识别出图 1 各组类簇，有效解决了任意形状、有噪声和非均匀类簇密度的聚类问题。

传统算法的粒度参数唯一性决定了不论粒度参数如何选择，它们都只能获取单一层面的聚类结果，因此无法全面反映数据的类簇关系和结构特征。在使用传统算法对图 1(d)中较大圆形区域的聚类特征识别中，聚合粒度取值过小时，上下两部分被分别识别为一个类簇；而聚合粒度取值

过大时,整个圆形被识别为一个类簇,这两种结果都是片面的。MSCA 的结果是一个关键聚合层的集合,根据这些聚合层则可以清楚地看到不同粒度下类簇之间的结构关系,得到数据全面且完整的内部结构和聚类特征。

4　结束语

针对非均匀类簇密度数据聚类问题,以商空间粒度计算理论和 CN-Tree 结构为基础,提出一种有效的多粒度自学习聚类算法 MSCA。MSCA 有效解决了传统算法中全局粒度参数无法全面地发现数据内部特征和粒度参数确定困难的问题。理论分析和实验结果表明,算法在发现任意形状类簇、有效处理噪声和异常点方面具有良好的特性,并能够准确发现关键聚合层,具有较好的聚类质量和效率,在数据多粒度分析和减小参数对经验的依赖性方面明显优于传统算法。

聚类分析已在动植物分类、疾病分类、图像处理、模式识别和文本检索等方面广泛应用,MSCA 以多粒度和自学习的方式将聚类分析范围扩展到非均匀类簇密度问题上,解决了粒度参数取值困难的问题,并使聚类分析的结果更客观、更全面,具有很强的实际应用价值。

参考文献:

[1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.

[2] MacQueen J. Some methods for classification and analysis of multivariate observations[C]// *Proc. of the 5th Berkeley Symposium on Mathematics Statistic Problem*, 1967:281-297.

[3] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// *Proc. of the Second International Conference on Knowledge Discovery and Data Mining*, 1996:226-231.

[4] Hinneburg A, Keim D. An efficient approach to clustering in large multimedia databases with noise[C]// *Proc. of the 4th International Conference on Knowledge Discovery and Data Mining*, 1998:58-65.

[5] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases[C]// *Proc. of the ACM SIGMOD*, 1996:103-114.

[6] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases[C]// *Proc. of the ACM SIGMOD International Conference on Management of Data*, 1998:73-84.

[7] Bezdek J C, Hathaway R J. Numerical convergence and interpretation of the fuzzy c-shells clustering algorithm[J]. *IEEE Trans. on Neural Networks*, 1992,3(5):787-793.

[8] Tari L, Baral C, Kim S. Fuzzy c-means clustering with prior biological knowledge[J]. *Journal of Biomedical Informatics*, 2009,42(1):74-81.

[9] Tang X Q, Zhu P, Cheng J X. Cluster analysis based on fuzzy quotient space[J]. *Journal of Software*, 2008, 19(4):861-868.

[10] 赵恒,杨万海. 基于属性加权的模糊 K-modes 聚类算法研究[J]. 系统工程与电子技术,2003,25(10):1299-1302. (Zhao Heng, Yang Wanhai. Fuzzy K-modes clustering algorithm based on the attributes weighted[J]. *Systems Engineering and Electronics*, 2003,25(10):1299-1302.)

[11] 厍向阳,薛惠锋,高新波. 基于障碍物约束的遗传-中心点聚类算法研究[J]. 系统工程与电子技术,2005, 27(10): 1803-1806. (She Xiangyang, Xue Huifeng, Gao Xinbo. Research on the genetic-medoid clustering algorithm with obstacles restriction[J]. *Systems Engineering and Electronics*, 2005,27(10):1803-1806.)

[12] 刘福才,马丽叶. 基于最邻近聚类和向量模糊 c-均值的混沌预测[J]. 系统工程与电子技术,2007,29(12):2162-2165. (Liu Fucai, Ma Liye. Prediction of chaos based on the nearest neighbor clustering and vector fuzzy c-means clustering[J]. *Systems Engineering and Electronics*, 2007,29(12):2162-2165.)

[13] 张钹,张铃. 问题求解的理论及应用[M]. 北京:清华大学出版社,1990.

[14] 张铃,张钹. 模糊商空间理论(模糊粒度计算方法)[J]. 软件学报,2003,14(4):770-776.

[15] Beckmann N, Kriegel H P, Schneider R, et al. The R*-tree: an efficient and robust access method for points and rectangles[C]// *Proc. of the ACM SIGMOD International Conference on Management of Data*, 1990:322-331.