文章编号：0258-2724(2010)02-0296-06 **DOI**：10.3969/j.issn.0258-2724.2010.02.023

基于灰关联测度的分裂式层次聚类算法

陈韬伟[1,2]， 金炜东[3]， 李 杰[2]

（1. 西南交通大学信息科学与技术学院，四川 成都 610031；2. 云南财经大学信息学院，云南 昆明 650221；3. 西南交通大学电气工程学院，四川 成都 610031）

摘 要：为估计数据集的聚类数目及获得较好的聚类性能，提出了一种基于灰关联测度的分裂式层次聚类算法. 该算法用灰关联测度衡量数据对象之间的相似程度，以基于密度扩展的方式自顶向下分裂成不同层次的数据集划分；然后，根据灰关联测度定义聚类有效性指标；最后将有效性指标曲线极值点对应的聚类划分用于估计最佳聚类数目. 实际数据和合成数据集的实验表明，与FCM聚类相比，该算法的聚类正确率平均提高3.7%，并且能够识别任意形状的簇.

关键词：灰关联测度；聚类分析；层次聚类；聚类有效性指标

中图分类号：TP31 **文献标识码**：A

Divisive Hierarchical Clustering Algorithm Based on Grey Relational Measure

CHEN Taowei[1,2]， *JIN Weidong*[3]， *LI Jie*[2]

（1. School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031, China; 2. Information College, Yunnan University of Finance and Economics, Kunming 650221, China; 3. School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China）

Abstract: To estimate cluster number and achieve a better clustering performance, a divisive hierarchical clustering algorithm based on grey relational measure was proposed. In this algorithm, the grey relational measure is used to measure the degree of similarity between data sets. On the basis of the way of density-based extension, the algorithm divisively generates hierarchical partitions of data set. And then the clustering validity index is defined based on the grey relational measure. The partitions corresponding to the extremum of the validity index curve are used to estimate the number of clusters finally. Computer simulation on real and synthesis data sets shows that compared with the FCM (fuzzy *C*-means) algorithm, the proposed algorithm has a 3.7% improvement in average clustering correct rate and is good for arbitrary-shaped clusters.

Key words: grey relational measure; clustering analysis; hierarchical clustering; clustering validity index

聚类算法广泛应用在统计、机器学习、模式识别、图像处理和数据分析等领域[1-3]. 目前已应用在多个领域内的聚类方法，大致可分为层次化聚类、基于密度和网格的聚类、划分式聚类. 但没有任何一种聚类算法可普遍适用于揭示各种多维数据集所表现出来的多样性结构[4]. 各种聚类算法都面临着样本间相似度的测量及聚类效果的评估问题，这些问题反映了算法设计者对类的定义和要求[5].

收稿日期：2008-09-16
基金项目：国家自然科学基金资助项目(60971103)
作者简介：陈韬伟(1976－)，男，博士生研究生，研究方向为智能信息处理和信号处理、雷达信号分选识别等，E-mail：cctw33@126.com
通讯作者：金炜东(1959－)，男，教授，博士，研究领域为智能信息处理、系统仿真与优化，E-mail：wdjin@swjtu.edu.cn

聚类算法通常是将样本之间的 Euclidean 或 Manhattan 距离作为相似性测量方法. 为了预先确定适合的聚类中心和最优聚类数目,文献[6-7]提出了利用灰关联分析(GRA)方法作为衡量样本之间联系的紧密程度,并结合划分式聚类提出了用有效性指标获得最佳聚类结果的评估. 但是在确定最终的聚类数和聚类结果时,需多次操作整个数据集,导致算法效率随数据集的增大而下降.

本文在此基础上,提出了基于灰关联分析的分裂式层次聚类 GRADHC (GRA-based divisive hierarchical clustering)算法. 它利用灰关联分析构成一种新的度量,结合分裂式层次聚类算法,对数据集逐层进行二分,直到叶子节点所代表的数据集满足不可分裂的条件为止,形成二叉树,并且一次性地由二叉树中不同层次的节点生成所有合理的划分,从而避免了对整个数据集的反复聚类. 然后根据新的有效性指标,构造不同层次划分的聚类质量曲线,该曲线的极大值点所对应的划分用来估计最佳的聚类数目.

在对每个节点的样本集进行二分时,采用了基于类似密度的聚类算法[8]实现样本集的分裂. 所以,GRADHC 算法能够识别数据集中可能包含的噪声和复杂形状簇. 针对真实数据和合成数据的实验结果,验证了所提算法的有效性和可行性.

1　相关概念

灰关联分析是通过灰色关联度来分析和确定系统诸因素的影响,或因素对系统主行为的贡献测度,其基本思路是根据对系统统计序列曲线几何形状的相似程度的比较,分清系统中多因素间的关联程度. 序列曲线的几何形状越接近,它们之间的关联度越大. 灰色关联分析已成功应用于决策、预测、控制、模式识别等方面[9].

由文献[9]可知,灰关联分析的主要步骤为:

(1) 确定参考序列和比较序列

选取数据集中的一个样本作为参考序列,记为 $x_i=\{x_i(k)\mid k=1,2,\cdots,d\}$, $i\in\{1,2,\cdots,n\}$,选取数据集中其余的样本数据作为比较序列,记为:

$y_{j\neq i}=\{y_j(k)\mid k=1,2,\cdots,d\}$, $j=1,2,\cdots,n$,

其中: d 是样本序列的维数; n 为数据集合中样本总数.

(2) 求关联系数

参考序列 $x_i(k)$ 与比较序列 $y_i(k)$ 的关联系数定义为:

$$\left.\begin{aligned} r_{ij}(k)&=\frac{\min\limits_j\min\limits_k\Delta_{ij}(k)+\xi\max\limits_j\max\limits_k\Delta_{ij}(k)}{\Delta_{ij}(k)+\xi\max\limits_j\max\limits_k\Delta_{ij}(k)},\\ \Delta_{ij}(k)&=|x_i(k)-y_i(k)|, \end{aligned}\right\} \tag{1}$$

其中: $\Delta_{ij}(k)$称为在第 k 个时刻 x_i 和 y_i 的绝对差; $\xi\in(0,1]$为分辨系数, ξ 越小,分辨力越大,一般取 $\xi=0.5$; $\min\limits_j\min\limits_k\Delta_{ij}(k)$ 称为两级最小差; $\max\limits_j\max\limits_k\Delta_{ij}(k)$称为两级最大差.

(3) 计算参考序列与各比较序列的灰关联度

将每一比较序列各个时刻的关联系数集中体现在一个值上以便于比较,这个值就是灰关联度,常用的计算灰关联度的方法是平均值法,即:

$$g(x_i,y_i)=\frac{1}{d}\sum_{k=1}^{d}r_{ij}(k), \tag{2}$$

式中, d 为数据序列的长度.

根据灰关联分析引入类似密度聚类的相关概念:

定义 1　对象 X 的近邻邻居:给定一个阈值 $T\geqslant 0$,数据集 D 中任意一个对象 X 的近邻邻居,记为 $N_T(X)$. 定义集合: $N_T(X)=\{Y\in D\backslash\{X\}\mid 1-g(X,Y)\leqslant T\}$,也称为对象 X 的最近邻集合,其中 $g(X,Y)$为 X 与 Y 的灰关联度.

定义 2　核心对象: D 为给定的数据集, $X\in D$,若 X 的近邻集合 $N_T(X)$在给定的阈值范围内非空,则称对象 X 为核心对象.

定义 3　核心集合:核心对象 X 的近邻邻居中,由所有核心对象加 X 本身构成的子集称为核心对象 X 的核心集合,记为 $E_T(X)$.

定义 4　对象 X 关于核心对象 Y 灰关联可达: $X,Y\in D$,对象 Y 为核心对象, $X\in N_T(Y)$,则称对象 X 关于核心对象 Y 的灰关联可达.

定义 5　核心对象 X 初始类:设 D 是数据集, $X\in D$,且 X 为核心对象, X 的初始类 C 是满足下列条件的数据集 D 的一个非空子集: $\forall X\in D$,若 $\exists Z\in E_T(X)$,使得 Y 关于核心对象 Z 的灰关联可达,那么 $Y\in C$.

显然,核心对象 X 的初始类 C 非空,且有 $N_T(X)\subseteq C$.

2　GRADHC 算法

层次化聚类方法是常用的聚类方法之一,可以进一步分为凝聚的和分裂的层次聚类. 目前的研究大多集中在层次凝聚算法[10],其基本思想是以数

据对象作为原子类,然后将这些原子类进行聚合,逐步聚合成越来越大的类,直至所有的个体都归为一类或达到某种可接受的标准为止. 而GRADHC算法则是自上而下的不断将数据集进行二分的过程,二分算法有很多选择,一般的方法是考虑所有的2^n-1种可能的划分,然后根据评价准则选择最优的一种划分. 这种方法虽然直观,但是计算量大,是一个NP-Hard问题. 为了避免上述问题,结合灰关联分析的思想提出数据对象的相似性测度的定义,以此为基础给出类似密度聚类算法来实现二分过程:

设d维数据集$D=\{X_1,X_2,\cdots,X_n\}$,n是数据集的数目,$X=\{x_1,x_2,\cdots,x_d\}$为其中的一个数据对象. 首先,将数据集D看成一类并设定初始阈值T,在包含有n个数据对象的数据集D中找到任意一个核心对象X,并求出X的核心集合,得到初始类;然后由初始类开始进行类的扩展,直至没有任何对象可以归入该类为止,扩展后的类记为C_{L};$|C_{\mathrm{L}}|$表示类C_{L}中包含的数据对象数目. 如果$|C_{\mathrm{L}}|<n$,数据集将分成两个类C_{L}和C_{R},当阈值T减少Δ时,如果$|C_{\mathrm{L}}|\leqslant|C_{\mathrm{R}}|$,则将$C_{\mathrm{R}}$再次分裂;否则继续对原数据集进行分裂,直到满足$|C_{\mathrm{L}}|\leqslant|C_{\mathrm{R}}|$为止. 这样递归地进行二分,每个阈值对应了一次分裂,将分裂后聚类树中各层叶子节点组成不同聚类数目的候选类,然后根据定义的有效性指标V_{G}来评估数据集的划分质量.

设g_{XY}表示核心对象X与比较对象Y的灰关联度,给定数据集D的一个划分$C=\{C_1,C_2,\cdots,C_k\}$,$S_{\mathrm{t}}(C)$衡量C的簇内紧凑度,$S_{\mathrm{p}}(C)$对应C的簇间分离度. 定义:

$$S_{\mathrm{t}}(C)=\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|C_i|\cdot|C_i|}\sum_{X,Y\in C_i}g_{XY},\tag{3}$$

$$S_{\mathrm{p}}(C)=\frac{1}{k(k-1)}\times\sum_{i=1}^{k}\left(\sum_{l=1,l\neq i}^{k}\frac{1}{|C_i|\cdot|C_l|}\sum_{X\in C_i,Y\in C_l}g_{XY}\right),\tag{4}$$

其中,$|C_i|$表示C_i包含的数据对象数目;S_{t}是簇内两个数据对象之间的平均灰关联度;S_{p}是分属不同簇的两个数据对象之间的平均灰关联度. 最优的聚类质量对应于簇内的紧凑度和簇间分离度的平衡点[11]. 另外,聚类的划分和阈值T有很大的关系,不同阈值下得到不同的聚类结果. 因此,算法使用的有效性指标$V_{\mathrm{G}}(C)$取以下形式:

$$\left.\begin{aligned}V_{\mathrm{G}}(C)&=Q(C)+\min\ T(C),\\Q(C)&=[S_{\mathrm{t}}(C)+S_{\mathrm{p}}(C)],\end{aligned}\right\}\tag{5}$$

其中,$\min\ T(C)$表示在不同阈值下出现相同划分的结果时选取最小的阈值. 根据灰关联分析可知,$S_{\mathrm{t}}(C)$的值越大,表明簇内对象间相似程度高,簇越紧凑;$S_{\mathrm{p}}(C)$的值越小,表明簇间的分离程度越好. 式(5)采用了线性组合来平衡$S_{\mathrm{t}}(C)$、$S_{\mathrm{p}}(C)$和$T(C)$的取值,最优聚类结果从数值上反映为指标函数$V_{\mathrm{G}}(C)$取得最大值.

总结以上所述,GRADHC算法步骤如下:

Step 1 初始化建立$n\times n$灰关联矩阵$\boldsymbol{M}=[M_{ij}]$,$M_{ij}=g(X_i,X_{j\neq i})$,$\boldsymbol{L}=1-\boldsymbol{M}$,将$\boldsymbol{L}$各列的元素按从小到大排序;初始化阈值$T=L_{n1}/z$,$z\in\mathbf{Z}^+$,$\Delta=L_{n1}/n$.

Step 2 将n个个体看成一类,即$C=\{X_1,X_2,\cdots,X_n\}$.

Step 3 以基于类似密度的扩展算法递归构建聚类二叉树.

Step 4 将分裂后保存的结果按不同层次组成不同聚类数目的候选类,通过矩阵$\boldsymbol{M}$计算每一T值对应的聚类有效性指标V_{G},选取满足最大V_{G}值对应的聚类数目以及聚类划分作为最终结果,算法结束.

算法过程中涉及到对阈值T和Δ的选取,根据定义1,T可以看作是聚类数据集的划分尺度,而尺度可以被想象为一个检测数据集中包含不同密度层次的"显微镜". 显然,Δ越小,分割成用于计算的区间数就越多,算法的时间开销也越大. 经过反复实验,选定Δ值,z选取2~5.

算法中,步骤1的时间复杂度为$O(n^2)$;步骤3是一个递归的过程,时间复杂度为$O(n\lg n)$. 故算法的时间复杂度为$O(n^2)$,与一般层次聚类算法时间复杂度相同[12-13].

3 仿真实验

仿真实验分别采用了真实的数据和人工合成的数据来评价算法的性能,并与几个典型的聚类算法进行了比较. 实验中用到的6个数据集:

(1) Iris:含有150个数据向量,形成3类,每个数据向量有4个属性,每类各有50个元素;其中,一类与另外两类线性可分,另外两类有部分重叠.

(2) Breast Cancer:Wisconsin breast cancer数据库含有699个数据向量,形成2类,每个数据向量有9个属性,每个类中的元素不同.

(3) Wine:含有178个数据向量,形成3类,每个数据向量有13个属性,每类中样本数量不同.

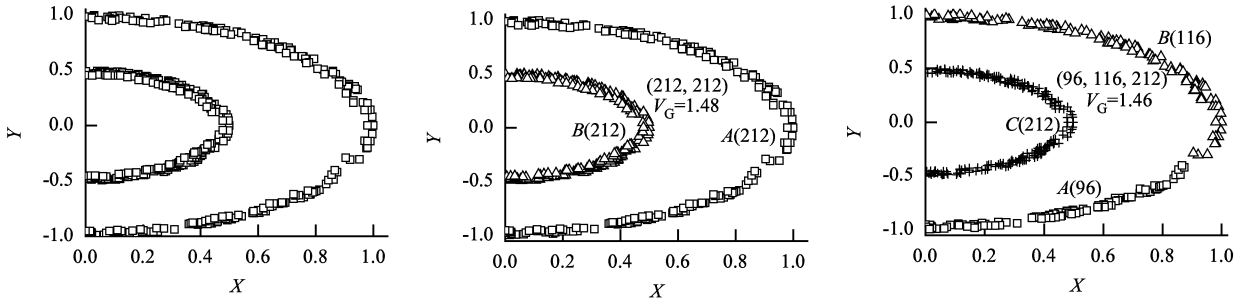
(4) G205_4k:含有 205 个数据向量,形成 4 类,每个数据向量有 80 个属性,类与类之间相距较近,其中两个类有较多元素,另两个类元素较少.

(5) 4k2bigsmall_lap:含有 400 个数据向量,形成 4 类,数目小的类靠近数目大的类,类之间有轻微重叠,数据向量有 2 个属性.

(6) Delta:含有 424 个数据向量,形成 2 类,具有特殊形状[14].

实验中,Iris、Breast Cancer 和 Wine 来自于 UCI 的机器学习数据库(http://www.ics.uci.edu/~mlearn/databases/),G205_4k 和 4k2bigsmall_lap 数据集均为人工数据集(http://www.mathworks.com/matlabcentral/fileexchage/).

为了更直观地反映算法的原理和性能,首先选取 Delta 数据集进行测试,图 1 给出了 GRADHC 算法在具有特殊形状数据集 Delta 上的若干分裂步骤和计算结果.



(a) 初始的簇　　(b) 分裂为二类及其有效性指标　　(c) 分裂为三类及其有效性指标

图 1　GRADHC 算法过程示例
Fig. 1　Clustering result for GRADHC

如初始状态图 1(a)所示. 假设所有数据点为一个独立的簇,随 T 的减小,数据集逐渐被分裂. 图 1(b)是在某阈值下分裂形成两个簇. 当阈值进一步减小时,数据集最终被分裂为图 1(c)的 3 个簇. 对于数据集在各层次上的划分结果(实际的聚类树在到达终止条件时可能不止 3 个层次,这里的例子假设只有 3 层),分别计算它们的有效性指标,抽取出使得指标值最大的划分,这样就得到了该数据集的最佳聚类数目为 2,最佳簇集合的划分如图 1(b)所示.

其次,选取真实 Iris 数据集来详细说明 GRADHC 的步骤:

从图 2 的二叉树中看出,当算法结束时,各叶子节点组成的划分在不同的阈值层次下构成不同的备选聚类:形成 2 类的划分为(50,100),形成 3 类的划分有(50,87,13)、(50,38,62)、(32,18,100),形成 4 类的有(50,38,33,29)、(32,18,38,62),形成五类的有(32,18,38,33,29),从而形成不同数目的划分类.

根据提出的有效性指标计算备选聚类的质量,表 1 列出了 Iris 数据集的 5 个备选聚类有效性指标. 从表中可以看出,算法在 Iris 数据集上可以得到正确的聚类数为 3,其最优的聚类结果为(50,38,62). 表 1 说明算法能够有效区分有重叠的簇.

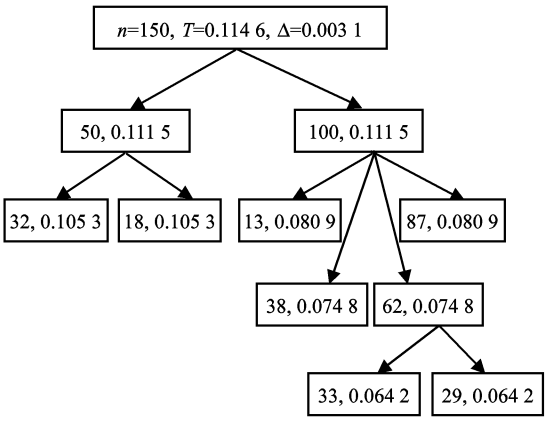


图 2　Iris 数据集分裂层次化聚类过程
Fig. 2　Process of divisive hierarchical clustering for Iris

表 1　Iris 数据集聚类有效性指标(V_G)
Tab. 1　Clustering validity index of Iris dataset

类别	C	V_G
2	(50,100)	1.574 6
3	(50,87,13)	1.531 2
	(50,38,62)	1.607 6
	(32,18,100)	1.586 4
4	(50,38,33,29)	1.556 6
	(32,18,38,62)	1.567 8
5	(32,18,38,33,29)	1.548 8

最后,针对其余 4 个数据集,利用 GRADHC 算法都得到了正确的最优聚类数. 对各数据集分裂的

部分划分结果由图 3 给出.

数据集在划分过程中,在不同的阈值下可能得到聚类数目相同的后备聚类,这是因为不同的划分尺度可以正确地反映簇之间的密度信息,如表 1 中的 Iris 数据集和 Breast Cancer 数据集的划分结果. 图 3(a)所示, Breast Cancer 数据集得到了最佳的聚类划分结果:(448, 251),其聚类准确率为 96.8%,高于其它一些算法的聚类结果. 此外, GRADHC 算法中的基于密度的划分也为快速得到最优聚类提供了保证,它通过数据集的本身结构进行扩展取得聚类数目,这正是算法能够在分裂层次较少的情况下获得正确聚类划分的原因,同时也是算法在识别特殊形状簇方面的优势.

表 2 列出了划分式聚类算法和层次聚类算法与 GRADHC 算法的聚类正确率(%)比较,其中括号中的数字表示各数据集的类别数量.

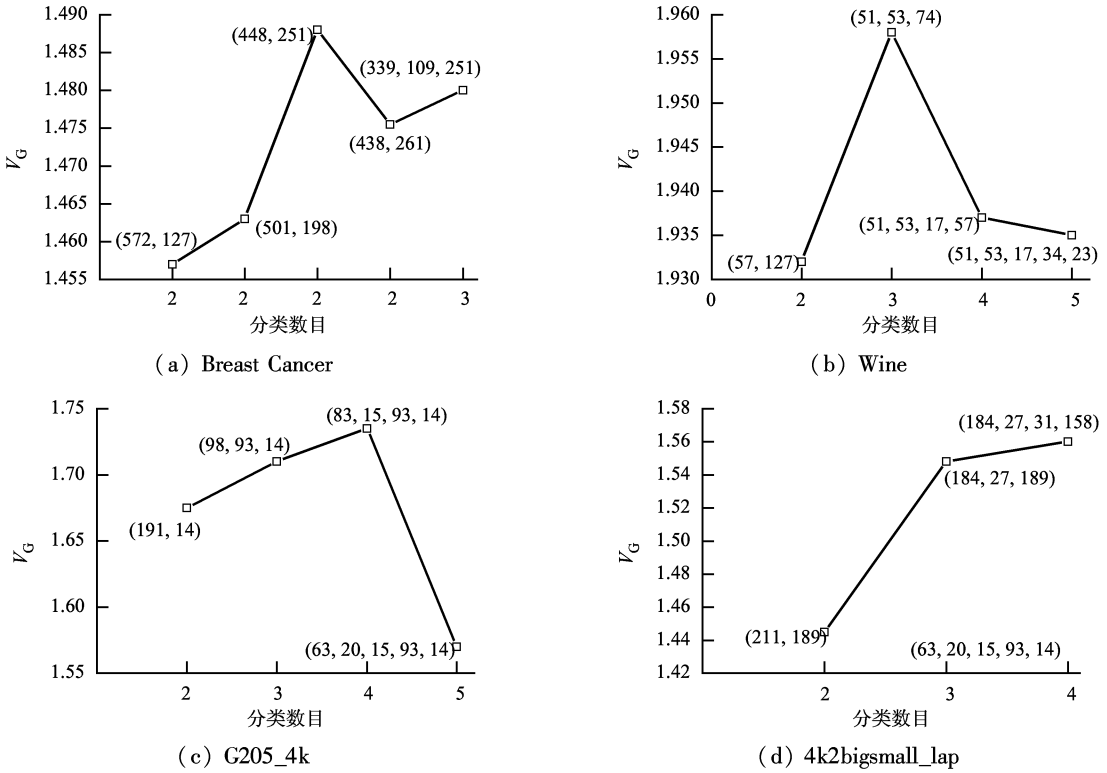


图 3 GRADHC 在 4 个数据集上的实验结果

Fig. 3 Experimental result for four datasets

表 2 不同算法对数据集的聚类正确率

Tab. 2 Comparison of clustering result for different algorithms %

算法	数据集/类别个数					
	Iris/3	Wine/3	Breast cancer/2	G205_4k/4	Delta/2	4k2bigsmall_lap/4
Fuzzy c-mean	89.33	68.54	95.27	83.41	50	95.0
K-mean	90.67	71.34	95.70	99.51	50	75.0
Single-link	68.00	42.70				
Complete-link	84.00	67.40	—	—	—	—
Group-average	74.70	61.20				
GRADHC	92.00	73.60	96.85	100	100	97.5

从表 2 中可以看出,传统层次算法中的单一联接法、完全联接法和类间平均联接法[15-16]对 Iris 和 Wine 数据集的分类结果不理想,这与传统层次聚类算法的再分配能力差有关. 此外,基于划分的聚类算法对特殊形状的数据集 Delta 分类效果不理想,这是因为划分聚类算法采用了半径的概念来控制聚类的边界,只能得到球形的聚类. 当采用 GRADHC 算法时,所有数据集均获得了较好的聚类结果,这是因为算法根据待聚类数据集的数据类型和聚类分布特点,自适应地取得最佳聚类的过

程.实验中也发现,划分聚类算法的运行效率要高于传统的层次聚类算法,但是随机选择初始中心导致了聚类的结果不稳定. GRADHC 算法的效率最低,但其聚类结果比较稳定,原因是算法以更多的时间开销为代价换取较高的分类准确性.

4　结　论

本文中提出了一种基于灰关联测度的分裂式层次聚类算法 GRADHC,用于获得最佳的聚类数目和分类结果.与目前大多数基于凝聚的层次聚类方法不同,它采用了自顶向下的密度扩展分裂过程,通过阈值 T 的逐渐减小发现不同层次的聚类结构,从而得到不同的备选分类结果.与其它聚类算法相比,它不需要多次运行特定的聚类算法,且从对象属性本身的几何形状衡量对象之间的关系.同时,算法使用新的有效性指标 $V_G(C)$ 的极值来估计聚类的个数,确定最佳分类结果.在文中不同数据集上的实验表明:算法在 Iris、Wine 数据集上的聚类正确率分别为 92% 和 73.6%,均高于其他聚类算法,并且在特殊形状的 Delta 数据集上获得了 100% 的聚类正确率.

算法中所涉及的参数选取对计算结果和效率将会产生影响.如何更好地确定参数、分析参数的影响以及算法对含有噪声的数据集上如何获得正确的聚类划分,将是下一步工作的重点.

参考文献:

[1] HANJ W, KAMBER M. Data mining: concepts and techniques[M]. San Francisco: Morgan Kaufmann Publishers, 2000: 335-391.

[2] 印桂生,于翔,宁慧. 基于粗约简的数据流增量聚类算法[J]. 西南交通大学学报,2009, 44(05): 637-643.
YIN Guisheng, YU Xiang, NING Hui. Incremental clustering algorithm based on rough reduction for data stream[J]. Journal of Southwest Jiaotong University, 2009, 44(5): 637 - 643.

[3] 胡学钢,曹永照,吴共庆. 一种有效的数据流二次聚类算法[J]. 西南交通大学学报, 2009,44(4): 490-494.
HU Xuegang, CAO Yongzhao, WU Gongqing. Effective twice-clustering algorithm for data streams[J]. Journal of Southwest Jiaotong University, 2009, 44(4): 490-494.

[4] SAMBASIVAM S, THEODOSOPOULOS N. Advanced data clustering methods of mining Web documents[J]. Issues in Informing Science and Information Technology, 2006(3): 563-579.

[5] 姜园,张朝阳,仇佩亮,等. 对聚类算法普遍存在问题的解决方法[J]. 电路与系统学报,2004,9(3): 92-99.
JIANG Yuan, ZHANG Zhaoyang, QIU Peiliang, et al. Solutions to general clustering algorithm issues [J]. Journal of Circuits and Systems, 2004, 9(3): 92-99.

[6] CHANG K C, YEH M F. Grey relational analysis based approach for data clustering[J]. IEE Proc.-Vis. Image Signal Process, 2005,152(2): 165-172.

[7] YEH M F, CHIANG S S. Grey ART network for data clustering[J]. Neurocomputing, 2005, 67: 313-320

[8] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proc. 2nd Int Conf on Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996: 226-231.

[9] DENG J L. Introduction to grey system theory[J]. J. Grey System, 1989, 1(1): 1-24.

[10] GRABMEIER J, RUDOLPH A. Techniques of clustering algorithms in data mining[J]. Data Mining and Knowledge Discovery, 2002, 6(4): 303-360.

[11] SUN Haojun, WANG Shengrui, JIANG Qingshan. FCM-based model selection algorithm for determining the number of cluster[J]. Pattern Recognition, 2004, 37(10): 2027-2037.

[12] JAIN A K, MURTY M N, FLYNN P J. Data clustering: a review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[13] GUHA S, RASTOGI R, SHIMK. Rock: A robust clustering algorithm for categorical attributes[J]. Information Systems, 2000, 25(5): 345-366.

[14] CAMASTRA F. A novel kernel method for clustering [J]. Pattern Analysis and Machine Intelligence, 2005, 27(5): 801-805.

[15] MARQUES J P. 模式识别——原理、方法及应用[M]. 吴逸飞,译. 北京: 清华大学出版社,2002: 51-74.

[16] FRED A L N, LEITÃO J M N. Partitional vs hierarchical clustering using a minimum grammar complexity approach [C/OL] // Proceedings of the SSPR&SPR 2000 [2008-10-29]. http://www.sigmod.org/dblp/db/conf/sspr/sspr2000.html.

(中文编辑:唐　晴　　英文编辑:付国彬)