聚类分析
在地球物理学研究中的应用

孙佳龙　沈立祥　李太春　著

WUHAN UNIVERSITY PRESS
武汉大学出版社

江苏省海洋技术品牌专业建设项目
江苏省海洋科学技术优势学科建设项目

聚类分析
在地球物理学研究中的应用

孙佳龙　沈立祥　李太春　著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

聚类分析在地球物理学研究中的应用/孙佳龙,沈立祥,李太春著.—武汉:武汉大学出版社,2018.12

ISBN 978-7-307-20640-3

Ⅰ.聚… Ⅱ.①孙… ②沈… ③李… Ⅲ.聚类分析—应用—地球物理学—研究 Ⅳ.P3

中国版本图书馆 CIP 数据核字(2018)第 266752 号

责任编辑:杨晓露 责任校对:汪欣怡 版式设计:马 佳

出版:**武汉大学出版社** (430072 武昌 珞珈山)

(电子邮件:cbs22@whu.edu.cn 网址:www.wdp.com.cn)

印刷:北京虎彩文化传播有限公司

开本:787×1092 1/16 印张:11.5 字数:273 千字 插页:1

版次:2018 年 12 月第 1 版 2018 年 12 月第 1 次印刷

ISBN 978-7-307-20640-3 定价:39.00 元

安徽大学图书馆 藏书 www.lib.ahu.edu.cn

版权所有,不得翻印;凡购买我社的图书,如有质量问题,请与当地图书销售部门联系调换。

目　录

此为试读,需要完整PDF请访问: www.ertongbook.com

第1章 绪 论

1.1 聚类分析基本原理

聚类分析起源于生物学的一个分支，生物学家为了研究生物的演变规律，根据各种生物的特征将它们归属于不同的界、门、纲、目、科、属、种之中。在考古学中，为了研究样品所属的年代，常常需要了解组成样品物质的特征，从而判别样品的归属。而在以往的分类学中，上述这些分类方法，主要是凭借经验和专业知识来进行的，很少与数学联系(方开泰等，1982)。由此可见，过去的分类，多数是按定性来进行的，很少用它们的特征数值定量地进行分类。然而，当样品的特征包含着诸多因素的影响时，单凭经验或专业知识来定性地分类是远远不够的。因此，利用数学方法进行定量地、科学地分类，已成为发展的必然趋势。

聚类分析是研究如何用数学的方法将事物进行分类的学科，其实质是将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程。聚类分析的目标就是在相似的基础上收集数据来分类。聚类源于很多领域，包括数学、计算机科学、统计学、生物学和经济学。在不同的应用领域，很多聚类技术都得到了发展，这些技术方法被用作描述数据，衡量不同数据源间的相似性，以及把数据源分类到不同的簇中。

假设我们根据 m 个指标的取值情况对 n 个对象 ω_1，ω_2，…，ω_n 进行分类，ω_i 的 m 个指标值放在一起记为 $(x_{i1}, x_{i2}, \cdots, x_{im})$，是一个 m 维欧几里得空间的点，$i=1, 2, \cdots, n$。聚类分析的基本原理是，按一定的方法定义 m 维欧几里得空间中的这 n 个点之间的距离，再利用这些距离的大小对这 n 个对象进行分类。

在聚类分析中，点 $(x_{i1}, x_{i2}, \cdots, x_{im})$ 和点 $(x_{j1}, x_{j2}, \cdots, x_{jm})$ 之间常用的距离有如下几种：

欧氏距离为：

$$d(i, j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \tag{1-1}$$

切比雪夫距离表示各坐标数值差的最大值，其公式为：

$$d(i, j) = \max_{1 \le k \le m} |x_{ik} - x_{jk}| \tag{1-2}$$

曼哈顿距离表示两个点在标准坐标系上的绝对轴距的总和，可用下式表示：

$$d(i, j) = \sqrt{\sum_{k=1}^{m} |x_{ik} - x_{jk}|} \tag{1-3}$$

对事物的分类指标定义之后就可以利用聚类方法对事物进行分类。主要的聚类算法可

以划分为如下几类：划分方法、层次方法、基于密度的方法、基于网格的方法以及基于模型的方法(Hong et al. , 2009；贺玲等, 2007；孙吉贵等, 2008)。

1. 基于划分的聚类方法

如果给定一个数据集 D，其包含有 n 个数据对象，用一个划分方法来构建数据的 k 个划分，每一个划分表示一个类，且 $k \leqslant n$。即它将数据对象划分为一个簇，并满足以下两点要求：①每一个簇至少包含一个数据对象；②每一个数据对象必须属于某一个簇。假定要构建的划分其数目为 k，划分方法是：首先，先创建一个初始的划分，然后，再采用一种迭代的重定位的技术，通过将数据对象在划分间来回地移动来改进划分(万志华等, 2005 ；郝洪星等, 2011；沈洁等, 2007；郭伟等, 2004)。一个好划分的准则为：同一类中的数据对象之间要尽可能的“接近”，而不同的类中的数据对象之间要尽可能的“远离”。

2. 基于层次的聚类方法

对给定的数据对象的集合进行层次的分解就是层次的方法(尉景辉等, 2005；彭京等, 2007；石剑飞等, 2008；曾志雄, 2007)。依据层次分解的形成过程，该方法可分为凝聚的层次聚类和分裂的层次聚类两类。自底向上进行的层次分解为凝聚的(agglomerative)层次聚类；自顶向下进行的层次分解为分裂的(divisive) 层次聚类。分裂的层次聚类先把全体对象放在一个类中，再将其渐渐地划分为越来越小的类，依此进行，一直到每一个对象能够自成一类。而凝聚的层次聚类则是先将每一个对象作为一个类，再将这些类逐渐地合并起来形成相对较大的类，依此进行，一直到所有的对象都在同一个类中方结束。

3. 基于密度的聚类方法

大多数的聚类方法都是用距离来描述数据间的相似性性质的，这些方法只能发现球状的类，而在其他形状的类上，这些方法都无计可施(许虎寅等, 2012；陈晋音等, 2015；李乐等, 2009；马帅等, 2003)。鉴于此，就只能用密度(密度实际就是对象或数据点的数目)将其相似性予以取代，该方法就是基于密度的聚类方法。密度的方法的思想是：一旦“领域”的密度超过某一个阈值，就将给定的簇继续增长。该方法还能有效地去除噪声。

4. 基于网格的聚类方法

基于网格的聚类方法是先把对象空间量化成有限数目的单元，将其形成一个网格空间，再对该空间进行聚类，这就是网格的方法(张伟莉等, 2008；邱保志等, 2006；程国庆等, 2009；刘敏娟等, 2007)。其主要优点为处理速度快，因为它的处理速度只与量化空间中的每一维的单元数目相关，而与数据对象的数目无关。

5. 基于模型的聚类方法

基于模型的聚类方法是先给每一个聚类假定一个模型，再去寻找能较好地满足该模型的数据的集合。此模型也许是数据点在空间中的密度分布的函数，也许是其他(张小涛等, 2008；魏瑾瑞, 2014；王利等, 2014；吴君浩等, 2007)。其潜在的假定为：一系列概率

的分布决定该目标数据的集合。统计方案、神经网络方案通常是其研究的两种方向。

1.2 聚类分析基本方法

1.2.1 基于划分的聚类算法

基于划分的聚类算法是给定一个包含有 n 个数据对象的数据库，用一个划分方法来构建数据的 k 个划分，每一个划分表示一个类($k \leqslant n$)。基于划分的聚类算法主要分为 K 均值(K-means)聚类算法、K 中心点(K-medoids)聚类算法、K 中值(K-median)聚类算法和 K 中心(K-center)聚类算法。

基于划分的聚类算法的缺点是要求事先给定聚类结果数，且对初始划分和输入顺序敏感等。为克服这些缺陷，万志华等(2005)以划分方法为基础，按密度从大到小，依距离选择较为分散的初始值，同时过滤噪声数据，并在聚类的过程中动态地改变聚类结果数，从而提出了一种基于划分的动态聚类算法，改善了聚类质量。苏锦旗等(2009)通过区域划分方法估算出 K 个中心点作为初始聚类中心，从初始聚类中心出发，应用 K-means 聚类算法，得到了聚类结果，从而提出了一种新的初始化 K-means 的聚类算法，结果表明，该算法能产生高质量的聚类结果、较少的迭代次数，优于 K-means 算法中传统的聚类中心初始化算法。

针对基于划分的经典聚类算法存在对初始聚类中心选取敏感的不足，孟娜娜等(2011)提出了一种基于划分的无监督优化算法。针对经典算法效率受噪声点影响以及对聚类研究对象形状存在局限的问题，引入层次聚类的思想，设计了基于遗传算法的初始聚类中心动态选取与优化聚类算法。实验结果表明，该算法可实现对任意形状、任意大小数据集快速有效的聚类。李玮等(2011)针对大数据集的初始聚类中心选取问题，在基于密度的划分算法和适用于大规模数据集限定初值的采样算法基础上，对聚类子空间在每一维上进行均匀划分形成不同的数据区域，根据数据区域的数据点数的多少进行采样来提高采样的准确性。利用采样思想缩小了数据集的规模，保证了算法在时间上的优势。通过不同规模、不同形状的数据集对算法进行验证，结果表明，该算法与其他初始聚类中心算法相比，在准确率和时间上都具有一定的优势。

赵恒等(2007)认为数据挖掘中解决分类属性数据聚类的算法有很多种，但大多数基于划分的方法得到的聚类中心一般不是数据集中的实际数据对象，缺乏实际的物理意义，有时会导致某一聚类为空。该文研究了近似 K-median 的求解算法，用数据的近似中值来代替模式进行聚类，提出了分类属性数据的近似 K-median 聚类算法，克服了一般基于划分的可分类属性数据聚类中所遇到的问题，仿真实验证明该算法有效。针对 K-means 算法存在依赖于初始聚类中心、容易陷入局部最优解等缺点，周浩理等(2015)提出了基于微正则退火 K-means 聚类算法。该算法通过继承微正则退火算法的高效全局寻优特性，避免了陷入局部最优解。结果表明，基于微正则退火 K-means 聚类算法能够有效地减少原算法对初始聚类中心点的依赖，从而提高算法的稳定性，摆脱了原算法容易陷入局部最优解的缺点。

针对 *K*-means 算法有对噪音数据鲁棒性不佳的不足，并对噪声和孤立点数据的敏感性，霍亮等(2013)将密度思想与 *K*-means 算法结合，提出了一种对 *K*-means 算法的改进算法，并通过实验表明了这种算法的可行性和有效性。王赛芳等(2011)针对传统的划分聚类算法不能够发现任意形状的簇的缺点，引入了一种能够有效反映样本间相似度的距离度量——基于路径的距离度量，同时，设计了一种能够反映类内样本相似度大、类间样本相似度小的目标准则函数。实验结果表明，基于路径划分的聚类算法与传统的 *K* 均值算法相比具有更好的聚类效果。针对 *K*-means 聚类算法容易陷入局部最优且对初始解很敏感的问题，党小超等(2010)提出了一种新的基于划分和连接度的聚类优化算法，明显地避免了对初始化选值敏感性的问题，并在 KDDCUP99 数据集中进行了检测，结果表明，该算法具有较高的检测率及较低的误检率。

针对 *K*-means 算法依赖于聚类中心初始位置的选择的不足，田生文等(2010)定义了多维属性对象的度、聚集度和聚集系数，选取度和聚集系数高的 *K* 个点作为 *K*-means 聚类的初始中心点，改进了 *K*-means 聚类算法，实验结果表明，改进后的 *K*-means 算法较传统的算法具有更高的效率和准确度。赵烨等(2012)提出了结合蚁群算法和 *K*-medoids 的聚类算法(AKCA)，解决了 *K*-means 算法需要人为确定簇数目，并对初始簇中心的依赖性较强的缺点，实验结果表明，AKCA 算法对于小型数据集具有运行效率高、聚类质量好和自适用性强等优点。

1.2.2 基于层次的聚类算法

基于层次的聚类方法的基本思想是：通过某种相似性测度计算节点之间的相似性，并按相似度由高到低排序，逐步重新连接各个节点。该方法的优点是可随时停止划分，主要步骤如下：

①移除网络中的所有边，得到有 *n* 个孤立节点的初始状态；

②计算网络中每对节点的相似度；

③根据相似度从强到弱连接相应节点对，形成树状图；

④根据实际需求横切树状图，获得社区结构。

在层次聚类算法中，一个分裂或合并被执行，就不能修正，因此，聚类质量受到限制。甄彤(2006)提出了利用簇间相异度及基于信息熵或整体相似度的聚类质量评价标准，在簇分裂过程中动态地进行簇的合并与分裂的算法。实验结果证明，该算法具有使结果簇更紧凑和独立的效果，具有更好的聚类质量。

胡建军等(2004)针对高维空间中任意形状的多层次聚类问题，基于“同类相近”的思想，提出并实现了最近邻优先吸收聚类算法 NNAF。实验结果表明，NNAF 算法适用于任意形状的高维空间数据的聚类，可以有效过滤噪声数据，对用户需要的先验知识少，可快速获得各种层次的聚类结果。

针对划分聚类对初始值较为敏感以及层次聚类时间复杂度高等缺陷，郝洪星等(2011)提出了一种基于划分和层次的混合动态聚类算法 HDC-PH。该算法首先使用划分聚类快速生成一定数量的子簇，然后以整体相似度的聚类质量评价标准来动态改变聚类数目，该算法的性能明显优于划分和层次算法，提高了聚类质量，并获得了更自然的聚类

结果。

曾志雄(2007)在综合分析基于划分的 K 均值聚类算法和基于层次的凝聚聚类算法的基础上，借鉴了各种混合聚类方法，提出了一种执行效率更高和聚类质量更好的分阶段混合聚类算法(HCAP)。通过在二维数据空间的模拟样本数据实验中验证，显示了该算法的有效性和合理性，在某些方面应用性能优于原算法。对于高维数据集，特别是对于数值属性的数据库或者分类属性的数据库的聚类，一直是一个难以解决的问题。董一鸿(2003)提出了一种基于邻域连接的层次聚类算法 HANL。该算法首先采用分割的方法将数据集划分为若干个子簇，通过对子簇间的连接的分析，建立子簇间的连接构成图，通过合并连接紧密度高的结点，便得到最后的聚类结果。该算法能够对任意形状的簇进行聚类，同时这种方法聚类速度快、效率高，具有良好的伸缩性，显示了该算法的有效性。

周晨曦等(2015)提出了一种基于约束动态更新的半监督层次聚类算法。该算法的优势在于省略了对必连约束的数据样本点进行预先划分的步骤，从而保证了数据样本点获得更为合理的聚合顺序，得到更为准确的聚类结果。实验表明，与其他同类算法相比，无论是在人工模拟数据集还是在现实数据集上，该算法都表现出了更高的准确性和更强的稳定性。叶茂等(2004)在层次聚类过程中，提出了一种新的层次聚类算法。该算法重新定义了簇与簇之间的距离度量，并以此为基础建立堆结构。利用估计数据点总体分布，证明了该算法将逼近最优解。实验结果也表明，该算法的聚类效果大大优于现有的聚类算法。

传统的层次聚类算法，每次迭代只将距离最小的那对类簇合并，容易受离群点影响，偏向于发现凸状或球状簇。受 CURE 算法启发，贾瑞玉等(2010)采用簇中固定数量代表点来代表簇对象进行距离的计算，并结合“90-10”规则，提出了一种改进的层次聚类算法 REPBFC(REpresentative Points Based Fast Clustering)，实验表明该算法是有效的。传统的凝聚层次聚类算法的时间复杂度为 $O(n^3)$，由于时间复杂度太高而无法应用到大的数据集。针对这一问题，姚玉钦等(2009)提出了一种新的基于网格的层次聚类算法。该算法先用基于网格的方法进行一次微聚类，然后再用凝聚的层次聚类算法进行聚类。在进行凝聚的层次聚类时，提出了一种新的簇间距离度量方法，该方法采用簇中权值最高的代表点的最小距离作为簇间的距离。理论分析和实验结果表明，基于网格的层次聚类算法比传统的凝聚层次聚类算法具有更高的效率和正确性。针对层次聚类算法一旦一个合并或分裂完成就不能撤销的缺点，李新良(2007)提出了一种新的层次聚类算法，即在每个组内进行原子聚类，从而得到原子簇集，然后将每个组内的邻近原子簇合并成子类，最后合并各分组内的邻近子类，得到最终的聚类结果，该算法具有使结果簇更紧凑和独立的效果，并且具有更高的效率。

1.2.3 基于密度的聚类算法

基于密度的聚类算法(DBSCAN，Density Based Spatial Clustering of Applications with Noise)主要的目标是寻找被低密度区域分离的高密度区域。基于密度的聚类算法可以发现任意形状的聚类，这对于带有噪音点的数据起着重要的作用。

在 DBSCAN 中，E 邻域是给定对象半径为 E 内的区域；如果给定对象 E 邻域内的样本点数大于等于 MinPts，则称该对象为核心对象；对于样本集合 D，如果样本点 q 在 p 的

E 邻域内，并且 p 为核心对象，那么对象 q 从对象 p 直接密度可达；对于样本集合 D，给定一串样本点 p_1，p_2，p_3，…，p_n，$p=p_1$，$q=p_n$，假如对象 p_i 从 p_{i-1} 直接密度可达，那么对象 q 从对象 p 密度可达；存在样本集合 D 中的一点 o，如果对象 o 到对象 p 和对象 q 都是密度可达的，那么 p 和 q 密度相连。

DBSCAN 算法需要用户输入 2 个参数：一个参数是半径(Eps)，表示以给定点 P 为中心的圆形邻域的范围；另一个参数是以点 P 为中心的邻域内最少点的数量(MinPts)。根据经验计算半径 Eps，然后根据得到的所有点的 k-距离集合 E，对集合 E 进行升序排序后得到 k-距离集合 E'，需要拟合一条排序后的 E' 集合中 k-距离的变化曲线图，然后绘出曲线，通过观察，将急剧发生变化的位置所对应的 k-距离的值，确定为半径 Eps 的值。如果觉得经验值聚类的结果不满意，可以适当调整 Eps 和 MinPts 的值，经过多次迭代计算对比，选择最合适的参数值。

基于密度的聚类算法具有挖掘任意形状聚类和处理“噪声”数据等优势，同时也存在时间消耗大、参数问题局限及输入顺序敏感等缺陷。为此，胡学钢等(2008)提出了一种基于层次树的密度聚类算法 DCHT(Density Clustering Based on Hierarchical Tree)，以层次树描述子聚类信息，动态调整密度参数，基于密度探测树结构中相邻子聚类得到最终的聚类簇。理论分析和实验结果表明，该算法适用于大规模、高维数据，并具有动态调整参数和屏蔽输入顺序敏感性的优点。以 DENCLUE 算法为基础，淦文燕等(2004)针对基于密度的聚类结果严重依赖于用户参数选择的缺点，提出了一种基于核密度估计的层次聚类算法。该算法首先优选窗宽 s 产生较好的核密度估计结果，然后以密度函数的局部极大值点为聚类中心形成数据的初始划分，最后根据密度函数的鞍点递归合并初始聚类产生不同层次的划分模式。仿真实验结果显示，该算法能够发现任意形状、大小和密度的聚类，并能够有效处理噪声数据，且聚类结果不依赖于用户参数的仔细选择。

陈治平等(2006)针对聚类中不规则形状数据点分布的处理难题，提出了一种基于密度梯度的聚类算法(CDG)。该算法通过分析数据样本及其周边的点密度变化情况，选择沿密度变化大的方向寻找不动点，获取了原始聚类中心。实验结果表明，新算法较基于密度的带噪声数据应用的空间聚类方法(DBSCAN)具有更好的聚类性能。为了减少 DBSCAN 聚类算法的查询次数，降低聚类的时间花费，石陆魁等(2005)在 DBSCAN 聚类算法的基础上，提出了一种基于密度的高效聚类算法。该算法首先对样本集按某一维排序，然后通过在核心点的邻域外按顺序选择未标记的样本点来扩展种子点，对样本进行非线性核变换后再进行聚类可以有效地改善聚类的质量。测试结果也表明新算法的时间复杂度和聚类质量都显著优于 DBSCAN 算法。

胡彩平等(2007)针对 DBSCAN 算法聚类时只考虑空间属性没有考虑非空间属性，且在对大规模空间数据库进行聚类分析时需要较大的内存支持和 I/O 消耗，他提出了一种改进的基于密度的抽样聚类算法，该算法不仅考虑了空间属性也考虑了非空间属性，从而能够有效地处理大规模空间数据库，在二维空间数据的测试结果表明，该算法是可行、有效的。由于基于密度的聚类算法不能自动处理密度分布不均匀的数据，崔尚卿等(2008)提出了一种基于不均匀密度的自动聚类算法。该算法既保持了一般基于密度算法的优点，也能有效地处理分布不均匀的数据。实验结果表明，该算法是有效的。

此为试读，需要完整PDF请访问：www.ertongbook.com

针对基于密度的聚类算法的聚类结果严重依赖于用户参数的选择，崔光照等(2006)将最小生成树理论与基于密度的算法相结合，提出了一种基于密度的最小生成树聚类算法。通过构造、分割最小生成树得到确定样本空间划分的最小生成子树，理论分析和应用结果表明，该算法不仅体现了基于密度聚类算法的优点，而且聚类结果不依赖于用户参数的选择，使数据聚类更合理，特别是对大型数据库非常有效。熊仕勇等(2011)针对网格和密度方法的聚类算法存在效率和质量问题，提出了密度与栅格相结合的聚类挖掘算法。该算法首先将数据空间划分为栅格单元，然后把数据存储到栅格单元中，利用 DBSCAN 密度聚类算法进行聚类挖掘，最后进行了聚类合并和噪声点消除，该算法在时间效率和聚类质量两方面都得到了提高。基于密度的聚类算法难以有效处理分布不均匀的线段对象集，康大伟等(2007)将 DBSCAN 中聚类的对象由点转变为线段，提出了一种基于密度的面向线段的聚类方法，从而发现了分布密度不同的各种簇，通过试验证明了该方法的可行性与有效性。

针对传统基于密度树网格聚类算法中存在人为设置密度阈值、重复查询邻域内对象以及边界点处理不当等问题，邢长征等(2016)提出了一种改进的基于密度与网格的聚类算法。首先将全部网格的平均密度值作为其密度阈值，避免了人为设置密度阈值的偏差；其次采用自适应算法确定密度半径，使其能适用到动态的聚类中；然后采用将邻域外未标记的点作为下一个核心点，依据分类情况进行扩展，对邻域对象的查询不再出现重复；最后对边界点进行了处理，增强了算法的聚类精度。实验结果表明，改进的算法在时间的效率及精度方面均有提高，并且能更好地适应聚类的动态性。高诗莹等(2017)针对 CFSFDP (Clustering by Fast Search and Find of Density Peaks)算法容易遗漏密度较小的类簇而影响聚类的准确率的缺点，提出了基于密度比例峰值聚类算法 R-CFSFDP。该算法将密度比例引入到 CFSFDP 中，通过计算样本数据的密度比例峰值来提高数据中密度较小类簇的辨识度，进而提升了整体聚类的准确率。

1.2.4 基于网格的聚类算法

基于网格的聚类方法将空间量化为有限数目的单元，形成一个网格结构，所有聚类都在网格上进行。基于网格的聚类方法采用空间驱动的方法，把嵌入空间划分成独立于输入对象分布的单元。基于网格的聚类算法使用一种多分辨率的网络数据结构。它将对象空间量化成有限数目的单元，这些网格形成了网格结构，所有的聚类结构都在该结构上进行(Kailing et al.，2003；Chu et al.，2009；Bouguessa et al.，2009)。这种方法的主要优点是处理速度快，其处理时间独立于数据对象数，而仅依赖于量化空间中的每一维的单元数。STING 算法和 CLIQUE 算法是常用的基于网格的聚类算法。

CLIQUE 算法是基于网格的空间聚类算法，但它同时也非常好地结合了基于密度的聚类算法，因此既能够发现任意形状的簇，又可以像基于网格的算法一样处理较大的多维数据。CLIQUE 算法把每个维划分成不重叠的社区，从而把数据对象的整个嵌入空间划分成单元，它使用一个密度阈值来识别稠密单元，一个单元是稠密的，即表示映射到它的对象超过密度阈值(Guo et al.，2003；Chairman Fayyad et al.，1999)。

STING 是一种基于网格的多分辨率的聚类算法，它将输入对象的空间区域划分成矩形

单元，空间可以用分层和递归方法进行划分。这种多层矩形单元对应不同的分辨率，并且形成了一个层次结构：每个高层单元被划分成低一层的单元。关于每个网格单元的属性的统计信息(如均值、最大值和最小值)被作为统计参数预先计算和存储。对于查询处理和其他数据分析任务，这些统计参数是有效的(Fraley et al.，1998；Avros et al.，2012)。

基于网格的聚类算法可以高效处理低维的海量数据，但对于维数较高的数据集，生成的单元数过多，导致算法的效率较低。刘俊岭等(2006)基于CD-Tree设计了新的基于网格的聚类算法，该算法在数据集的大小及维度的可伸缩性方面均有显著提高，效率远高于传统的基于网格聚类算法。

张伟莉等(2008)针对基于网格的聚类算法所需的初始参数比较复杂的缺点，提出了一种新的基于网格的聚类算法CABG。该算法利用网格处理技术对数据进行了预处理，能根据数据分布情况动态计算每个单元格的半径，并成功地将网格预处理后所得单元格数据运用于其后的聚类分析中，从而简化了算法所需的初始参数。实验表明，CABG算法不仅具有DBSCAN算法准确挖掘各种形状的聚类和很好的噪声处理能力的优点，而且具有较高聚类速度以及对初始参数较低的敏感度。陈卓等(2005)针对传统网格聚类算法聚类质量较低的缺点，提出了快速聚类算法。通过凝聚点来准确反映数据空间的几何特征，然后采用网格和密度相结合的方法，利用爬山法和连通性原理进行了聚类处理。实验结果证明，该算法的聚类效率优于传统爬山法、Clique算法和DBSCAN算法。

为了提高基于网格技术的聚类精度，邱保志等(2006)提出了利用低密度单元中的点到高密度单元中心的距离作为判断聚类边界点和孤立点的技术，开发了HQGC算法。实验表明，该算法能识别任意形状的聚类，聚类的精度高、运行速度快、可扩展性好。单世民等(2008)研究发现现有的基于网格聚类算法在付出较小的时间复杂度的同时，牺牲了聚类的质量，得到的往往并不是最理想的聚类结果，尤其是在簇边缘可能出现数据点聚类不准现象，他提出了一种将网格化空间中位于簇边缘的网格进行精度进一步细化处理的算法，将这些边缘网格中的这些不确定的点重新恢复它们的固有信息，再利用相似度函数将它们分配到合适的簇中。在空间数据集上的实验数据表明，这种簇边缘精度增强聚类算法可在$O(n)$时间复杂度内得到优于CLIQUE算法的聚类结果。

余灿玲等(2010)通过研究发现，现有的基于网格的聚类算法在获得较高效率的同时，却是以牺牲聚类的质量为代价的，特别是在簇与簇相互邻近的情况下。为此，她提出了一种基于网格密度方向的聚类预处理方法。当一个网格单元密度出现反方向递增时，即挤压的情况，则需要对该单元进行进一步的细分处理，判断该单元是不是簇的边缘单元，并准确地判断边缘单元中对象的挤压方向。实验显示该算法可以有效地加强聚类簇边缘的精度，具有较高的簇识别率，因此，作为聚类的预处理算法是理想的。针对传统网格聚类算法聚类质量较低的缺点，梁敏君等(2009)提出了一种基于网格和分形维数的聚类算法，该算法结合了网格聚类和分形聚类的优点，改进了分形聚类耗时较大的问题。即首先根据网格密度得到初始类别，再利用分形的思想，将未被划分的网格依次归类。实验结果证明，该算法能够发现任意形状且距离非邻近的聚类，且适用于海量、高维数据。

邱保志等(2009)为了解决相交网格划分技术中聚类结果对数据输入顺序的依赖性和聚类结果精度不高的问题，提出了一种基于相交划分的动态网格聚类算法DGBO。该算法

利用相交网格划分技术和移动网格技术来解决上述问题，通过连接相交的高密度网格单元形成聚类，只需一个参数，运行速度快。实验表明，DGBO 算法能够快速有效地对任意形状、大小的数据集进行聚类，并能很好地识别出孤立点和噪声。邱保志等(2007)为了解决动态网格划分技术中聚类结果对数据输入顺序的依赖性和聚类精度差的问题，提出了基于移动网格技术的动态网格聚类算法。该聚类算法主要利用了动态网格划分技术和移动网格技术来解决上述问题，且能够识别任意形状、任意大小的聚类，只需一个参数，且时间复杂度是数据集大小和数据维度的线性函数，实验结果表明该算法是有效的。而针对网格聚类算法中的输入参数和聚类结果不精确问题，邱保志等(2010)提出了基于局部密度的动态生成网格聚类算法 DGLD。该算法使用动态生成网格技术能大幅度地减少数据空间中生成的网格单元的数量，并简化邻居的搜索过程；采用局部密度思想解决数据空间相邻部分对网格密度的影响，提高了聚类精度。

1.2.5 基于模型的聚类算法

基于模型的聚类方法的基本思想是：为每个聚类假设一个模型，再去发现符合模型的数据对象，试图将给定数据与某个数学模型达成最佳拟合。一个基于模型的算法可能通过构建反映数据点空间分布的密度函数来定位聚类，也可能基于标准的统计数字自动决定聚类的数目，同时考虑“噪声”数据和孤立点，从而产生健壮的聚类方法。这种聚类方法总是试图优化给定的数据和某些数学模型之间的适应性(宋浩远等，2008)。典型的算法主要有 Mrkd-trees 算法、粒子筛选(Particle Filters)、SOON 算法和混合算法。

Mrkd-trees 算法是基于 mrkd(multiple-resolution k-dimension)结构的一棵包含一定数量信息的结点构成的二叉树(Doucet et al.，2001)。树中的结点分为叶结点或非叶结点。树中的每个结点存储的信息为超矩形的边界和统计量集。粒子筛选是一种把蒙特卡罗方法应用于动态状态-空间系统的序列方法(Xu et al.，1998)。粒子过滤器用 N 个加权粒子来评估一些兴趣量。如果在 t 时刻的模型的状态-空间被表示为 s_t，那么粒子 i 就是一个特殊状态即 $s_t^{(i)}$ 的展现。每个粒子被赋予一个权值即 $\omega_t^{(i)}$。因此，一个粒子过滤器可以看作粒子和权值之集即 $\{s_t^{(i)}, \omega_t^{(i)}\}_{i=1}^{N}$。SOON(Self Organizing Oscillator Networks)算法利用神经网络把对象集组织成 k 个稳定而被结构化的簇(Rhouma et al.，2001)。k 值以无监督的方式确定。SOON 算法的基本思想是把空间上大量相关的代表对象放置到一维或二维空间，彼此邻近的代表对象一起被看作相似(Xu et al.，1998)。每当一个观测值出现，最邻近的代表对象也被发现，同时更新所有代表对象的值；数据不断被提交直到收敛。混合算法是集基于模型、密度和网格的方法于一体的一种方法(Rhouma et al.，2001；Doucet et al.，2001)。该方法的主要思想是确定簇的核心是基于密度的方法，即与其他区域相比，样本空间中的簇显得更稠密。稠密区域的显著特征是域内点的最邻近距离要小于域外点的最邻近距离。

现有的聚类算法一般只能处理以固定间隔表示的数据类型，而忽略了时间轴的变化。张小涛等(2008)提出了基于倒谱距离测度和自回归条件持续期(ACD)模型的聚类方法。该方法综合了计量模型的参数估计和聚类的非参无监督分类的优点，是一种适合处理不等间隔时间序列的技术。实验结果证明，这种方法是有效的，从中得出的结论与市场微观结

构理论也是相吻合的。

在聚类问题中，若变量之间存在相关性，传统的应对方法主要是考虑采用马氏距离、主成分聚类等方法，但其可操作性或可解释性较差，因此，魏瑾瑞(2014)提出了一类基于模型的聚类方法。该方法先对变量间的相关性结构建模(作为辅助信息)再做聚类分析。这种方法的优点主要在于：适用范围更宽泛，不仅能处理(线性)相关问题，而且还可以处理变量间存在的其他复杂结构生成的数据聚类问题，而各个变量的重要性也可以通过模型的回归系数来体现，因此，比马氏距离更稳健、更具操作性，比主成分聚类更容易得到解释，算法上也更为简洁有效。

传统的聚类算法如 *K*-means 算法需要一些先验知识来确定初始参数，初始参数的选择通常会对聚类结果产生很大的影响。王维彬等 (2007)提出了一种新的基于模型的聚类算法，通过优化给定的数据和数学模型之间的适应性发现数据对模型的最好匹配。高斯混合模型可以看作一种“软分配聚类”方法，该方法结合一种贪心的 EM 算法来学习高斯混合模型(GMM)，由贪心 EM 算法实现高斯混合模型结构和参数的自动学习，而不需要先验知识。因此，这种聚类算法可以克服 *K*-means 等算法的缺点。实验结果表明该算法具有更好的聚类效果。与传统的硬划分聚类相比，模糊聚类算法(以 FCM 为例)对数据的比例变化具有鲁棒性，能够更准确地反映数据点与类中心的实际关系，目前已得到广泛应用。然而对于时序基因表达数据来说，传统的聚类算法往往不能充分利用到数据中时间上的动态关联信息。因此，刘宇宏等(2008)在模糊聚类算法的基础上引入自回归(AR)模型，将时序基因表达数据作为一组时间序列进行动态的聚类分析。这样不仅可以充分利用到时序基因表达数据的内部自相关性，并且可以进一步利用隶属度函数对 AR 模型的预测过程进行模糊化调整，从而得到更为理想的聚类结果。

针对已有基于模型的多维时间序列(MTS)聚类算法处理不等长 MTS 速度较慢的问题，霍纬纲等(2017)提出了一种基于 LR 分量提取的 MTS 聚类算法 MUTSCA。该算法采用等频离散化方法符号化 MTS，计算用于表达 MTS 样本各维时间序列之间时序模式的 LR 向量，对每个 LR 向量进行排序后从其两端提取固定数目的不同关键分量，最后，采用 *K*-means 算法对生成的等长模型向量集进行聚类分析。实验结果表明，该算法能够在保证聚类效果的前提下，显著提高不等长 MTS 数据集的聚类速度。鉴于传统聚类算法需要参数假设、限于局部最优等不足，何会民(2008)假设数据点之间存在随机游走关系，根据数据相似性构造随机游走过程的转移矩阵，当随机游走过程进入收敛期后，t 阶转移矩阵揭示了数据点的分布，用迭代方法寻找最小的 KL-divergence 来对这些分布聚类。实验表明，该算法具有较优的聚类效果。传统的 *K*-means 算法由于其方法简单，在模式识别和机器学习中被广泛讨论和应用。但由于 *K*-means 算法随机选择初始聚类中心，而初始聚类中心的选择对最终的聚类结果有着直接的影响，因此算法不能保证得到一个唯一的聚类结果。曹付元(2008)利用邻域模型中对象邻域的上下近似，定义了对象邻域耦合度和分离度的概念，给出了对象在初始聚类中心选择中的重要性，进而提出了一种初始聚类中心的选择算法。通过与随机选择初始聚类中心、CCIA 选择初始聚类中心算法进行比较，实验结果表明，该算法是有效的。

1.3 聚类分析基本应用

1.3.1 聚类分析在生物学中的应用

张焱等(1999)根据14个定量性征指标的数据，应用5种系统聚类方法，对采自四川境内的若干金丝猴头骨进行了聚类分析，得出了基本一致的分类结论，根据这个分类结果，确定了三例产地不详的金丝猴头骨的原产地，并探讨了川金丝猴的地理演化。

红色糖多孢菌的摇瓶培养基中使用不同有机氮源时红霉素的效价明显不同。罗致强等(2006)通过对培养基中不同有机氮源所含无机盐进行了聚类分析。研究发现无机磷为影响红霉素效价的主要因素。

在分析了杭州郊区秋末冬初季节小白菜上菜蚜种群的数量动态和空间动态的基础上，刘树生等(1994)应用模糊聚类分析法研究了菜蚜种群的数量和空间的整体动态。结果表明，桃蚜、萝卜蚜及其混合种群的数量动态均呈指数或Logistic曲线变化，它们的空间格局呈聚集分布；而且聚集强度始终从高到低呈持续下降。运用模糊聚类法，可将其种群的整体动态分成4个时期，即苗期(或移栽后的返青期)的迁入期，成株初期的增殖扩散期，成株后生长盛期的繁殖高峰期，外围叶片明显枯黄时的数量饱和期。

卜耀军等(2005)应用模糊聚类分析方法对黄土高原丘陵沟壑区植物群落演替进行研究，研究结果表明，利用模糊聚类分析，可以将13个样地客观地划分为5个群系。该演替序列与传统的研究结果基本相同，说明利用模糊聚类分析方法研究植被演替是可行的。

郭水良等(1995)以不同山地的植物区系为比较对象，以地区植物区系中属的地理成分为指标，应用模糊聚类分析中的最大树法，比较了全国16个山地的植物区系关系。结果表明，应用模糊聚类最大树法得到的16个地区植物区系关系的最大树，能比较客观地体现出这16个地区之间植物区系关系，而且图形直观，方法简便，具应用价值。

王显金等(2011)从DNA序列片段个案中密码子分布密度角度出发，提取出DNA序列片段的特征，基于五大类氨基酸出现的频率，应用聚类分析方法对DNA序列片段进行分类。结果表明，该算法具有分类简单且分类结果精度较高的优点。

黄天带等(2012)分析了16种不同基因型橡胶树幼果内珠被愈伤组织诱导率、分化率，基于25对EST-SSRs分子标记数据对其进行聚类分析，研究系谱、聚类结果与体胚发生的关系。体胚发生结果与系谱分析结果基本吻合，与聚类分析结果吻合度不大，提示可以利用系谱分析提高橡胶树组织培养的培养基筛选效率。

1.3.2 聚类分析在信息与通信工程中的应用

为帮助盲人更好地利用盲道，闫志杰等(2010)利用颜色聚类分析的方法对图像进行初步的区域分割，然后根据盲道的颜色特征从中选择属于盲道的区域。利用拉东变换对图像中的直线边缘进行检测，并结合对盲道的初步分割结果找到盲道的边缘，以实现对盲道区域的精确分割，进而提出了一种自适应盲道分割算法。通过在不同时间和不同的天气条件下采集室外的盲道图像对算法进行测试，实验表明，环境变化对该算法影响很小，可以

实现对盲道区域的自适应分割。电信业正面临经营环境和市场格局的一系列变化。电信运营商通过培育企业的核心竞争力，采用“客户为中心”的CRM管理模式进行客户关系管理，了解顾客的消费模式并向其提供满意的产品和服务，企业才能生存和发展。郑国荣等(2006)通过比较常用的几种聚类算法，提出了改进的算法并应用于客户的消费模式分析，得到了较好的效果。

针对光网络中的多故障定位问题，宋继恩等(2013)提出了光网络中多故障定位的模糊聚类模型。根据由模糊聚类算法得到的专家系统和由此确定的告警和故障之间的隶属函数公式进行故障定位。在格状光网络拓扑下进行多故障定位仿真实验，得到了故障定位成功率的对比图和时间性能的对比图。实验结果表明，模糊聚类算法可以高效快速地定位故障。

在网络安全事件中大部分为异常事件，刘鹏等(2013)把聚类分析和孤立点技术引入网络安全态势评估，提出了一种利用聚类数据集合和孤立点数据集合计算服务层威胁值的方法，为评价网络安全态势提供了一个可以参考和决策的重要参数。

针对电信运营企业的仓库布局的主要问题，祝丽等(2007)应用聚类分析方法，提出了仓库布局优化方案，并进行了方案的综合评价，得到了良好的效果。

随着信息技术的高速发展，聚类分析方法在客户分类中得到越来越多的应用。施卓敏等(2014)借助SPSS Clementine 12.0软件，使用两步聚类分析方法对中国科学院资源规划项目(Academia Resource Planning，ARP)系统用户进行聚类分析，总结了各类用户的特点，剖析聚类结果的原因，并针对不同类型用户存在的问题提出了具体的解决建议，为提升系统使用效果和服务水平提供参考依据。

社会媒体网络属于典型的复杂网络，黄蓝会(2016)将复杂网络中的聚类分析方法应用到社会媒体网络中的社团结构研究，从而提出了将网络中的节点转化成代数空间上的数据对象，通过节点相似性系数来研究社团结构，最后通过仿真实验证明，利用节点相似性系数得到的社团发现准确率较高。

按照自动售取票旅客特征分别提供服务是提高铁路自动售取票工作效率的有效手段。为此，车站需要在不同的区域安装不同类型的自动售取票终端，并配置不同的操作界面。郭畅(2017)在借鉴互联网用户行为分析技术的基础上，提出了一种适应于铁路自动售取票旅客特征分析的技术，包括*K*-means聚类分析和Top N分析。应用这些分析技术能够为车站提供自动售取票旅客的聚类特征和最常用的售取票方式，从而可以指导车站进行自动售票终端的布局和终端软件界面配置。

蔡洪民等(2016)基于Wireshark进行IPv6数据包的捕获解析并存储，然后使用MATLAB聚类工具箱中的*K*均值算法和神经网络工具箱中的SOM算法，分别对包含多类攻击数据的IPv6流量进行处理，从而实现了对于CERNET2网络的异常流量聚类识别。实验表明，该系统能够识别发生在IPv6网络中的DOS攻击等几类针对ICMPv6的攻击，加强了校园网络的安全。

1.3.3 聚类分析在地质资源与地质工程中的应用

吕红华等(2006)利用244口井的小层数据，用Q型主因子分析与聚类分析相结合的

此为试读，需要完整PDF请访问：www.ertongbook.com