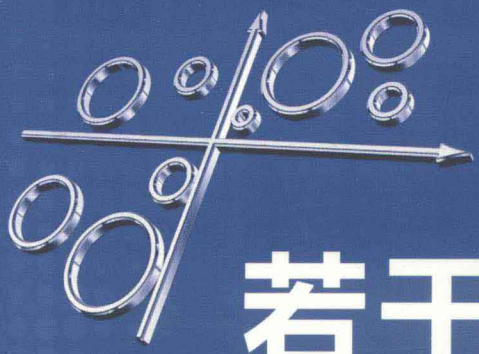


若干聚类问题复杂性及其算法

◆ 刘培强　李曙光　肖进杰　著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

若干聚类问题复杂性及其算法

刘培强　李曙光　肖进杰　著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

聚类是指根据给定的多个对象及其属性，基于相似性函数度量对象间的相似性，以寻找有意义或有用的对象分组。聚类分析方法是人们认识和理解世界的最基本方式之一，广泛应用于计算生物学、市场分析、社交网络数据分析、电子商务数据分析等众多领域。由于聚类分析的多样性、重要性和广泛性，尤其是在目前大数据时代背景下，众多应用领域对聚类分析算法提出了新的挑战。本书从问题的计算复杂性证明和近似算法设计的角度，对若干个聚类问题进行了讨论和研究，主要研究了带缺失值的两元指纹向量聚类问题、两元矩阵的 k-子矩阵划分问题、割聚类问题、设施定位问题与 k-median 问题等。本书可作为从事计算复杂性理论、聚类分析研究和应用科技人员的参考书。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

若干聚类问题复杂性及其算法 / 刘培强，李曙光，肖进杰著. —北京：电子工业出版社，2013.10
ISBN 978-7-121-21379-3

Ⅰ. ①若… Ⅱ. ①刘… ②李… ③肖… Ⅲ. ①基因表达—聚类分析 Ⅳ. ①Q786

中国版本图书馆 CIP 数据核字（2013）第 209124 号

策划编辑：薄 宇
责任编辑：底 波
印 刷：三河市鑫金马印装有限公司
装 订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：720×1 000 1/16 印张：9.25 字数：151 千字
印 次：2013 年 10 月第 1 次印刷
定 价：35.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

聚类分析方法是人们认识和理解世界的最基本方式之一，是我们获取知识的重要方法，它广泛应用于计算生物学、金融数据分析、市场分析、社交网络数据分析、电子商务数据分析等众多领域。由于聚类分析的多样性、重要性和广泛性，尤其是在目前大数据时代背景下，众多应用领域为聚类分析算法提出了新的挑战，使得有60余年的聚类分析方法仍是目前研究的热点问题之一。

聚类分析是指根据给定的多个对象及其属性，以对象属性为参数设计相似性函数，基于相似性函数度量对象间的相似性，以寻找有意义或有用的分组，具有相似特征的一组对象称为一个簇（Cluster）。若相似性函数的参数为对象的全部属性，则称此种聚类为单向聚类；若相似性函数的参数为对象全部属性的一个子集，则称此种聚类为双向聚类。

本书从计算复杂性的角度对若干个聚类问题进行了讨论和研究，这些聚类问题包括：

（1）带缺失值的两元指纹向量聚类问题；

（2）两元矩阵的 k-子矩阵划分问题；

（3）割聚类问题；

（4）设施定位问题与 k-median 问题。

本书共分为7章。第1章为绪论，介绍聚类的基本概念和本书所讨论的聚类问题。第2章介绍计算复杂性理论。第3章讨论基因表达谱分析的两元指纹向量聚类问题的复杂性，给出了两个求解算法。第4章讨论两元矩阵的 k-子矩阵划分问题（k 是正整数常量，下同）和二分图的二分团划分问题；先后证明了 MO3 问题的一个变体形式、3-二分团划分问题、k-二分团划分问题（$k>3$）、k-子矩阵划分问题（$k>3$）是 NP-完全的，并给出了二分图

的 k-二分团划分问题的一个指数精确算法。第 5 章证明链和环上最优聚类的一个良好结构，并据此结构设计了一个最优算法，然后为树和限制树宽图中的这一问题设计一个两阶段算法，其近似性能比为 $2-1/2k^2$，其中 k 表示给定终端的数目。第 6 章讨论颜色相关最小负载聚类问题，给出固定参数可解算法。第 7 章首先分析设施定位问题局部搜索求解算法的求解质量，并编程测试对比贪心算法和随机算法；然后提出一个解决 k-median 问题的贪心算法，实验证明此算法实用效果较好。

本书由刘培强、李曙光、肖进杰著，第 1、2、3、4 章由刘培强编写，第 5、6 章由李曙光编写，第 7 章由肖进杰编写。

本书的出版工作得到了国家自然基金（61173173）、教育部科学技术研究重点项目（2012101）、山东省自然科学基金（ZR2011FL004, ZR2011FM035）、烟台市科技发展计划项目（2010167）、山东省高等学校科技计划项目（J11LG14）、山东省科学技术发展计划（软科学）（2013RKB01127）等项目和山东省高校智能信息处理重点实验室（山东工商学院）的资助，在此表示感谢！

此外，还要特别感谢山东大学的朱大铭教授，山东工商学院的范辉教授、原达教授、谢青松教授，感谢他们为本书写作提供的帮助与支持。

本书可作为从事计算复杂性理论、聚类分析研究和应用科技人员的参考书。

由于我们学术水平有限，书中难免有错误和疏漏之处，敬请读者指正。作者的联系方式：liupq@126.com。

作　者

2013 年 6 月

此为试读，需要完整PDF请访问：www.ertongbook.com

目 录

第1章 绪 论

1.1 聚类分析

在人们认识、理解世界时，“类”的概念具有非常重要的意义，如在生物学中，人们为了研究生物，将所有的生物体分为域、界、门、纲、目、科、种。类是具有相似公共属性的一组对象。将多个对象分到不同类中的主要技术是分类和聚类。简单地讲，分类是指按照事先给定的目标函数，赋予对象不同的类标号，从而将对象分到不同类中；聚类是根据给定的多个对象及属性，基于对象的属性设计相似性函数，由相似性函数度量对象间的相似性，在对象中寻找有意义或有用的分组，具有相似特征的一组对象称为一个簇（Cluster）[①]。分类与聚类的主要不同点在于，分类是一种监督分类（Supervised Classification）技术，而聚类是一种非监督分类（Unsupervised Classification）技术。

聚类分析的目的主要包括：

（1）获取隐藏在数据中的自然结构；

（2）获取对象的自然分类；

（3）数据压缩。

如果聚类的目的是为了理解对象，则可以将簇看作潜在的类，而聚类技术就是自动寻找簇的一种技术。例如，在分子生物学中，使用聚类技术对基因表达谱进行分析，结果可用于推导基因的功能。如果聚类的目的是为了实

① 也有文献将聚类结果“Cluster”译为“类”。

用，则聚类是为了寻找刻画簇特征的簇原型，而聚类技术就是寻找最有代表性簇原型的技术。例如，使用聚类技术寻找对象数据中的簇原型，并建立簇原型表，然后将每个对象用其在簇原型表中的索引号表示，从而达到数据压缩的目的[1]。

就聚类算法本身而言，经过近 60 年的发展，已经广泛应用于众多领域。无论是聚类应用领域还是聚类算法，都是数以千计的。在 Google 学术搜索（2013 年 5 月）中，以“clustering”作为关键词进行搜索，可以找到 220 万余条目，2010 年以后有 33 万余条目，这说明了聚类分析在数据分析中的重要性和普遍性。现在，随着大数据时代的来临，在基因表达谱分析、物联网、互联网、信息检索、商业等领域中，给聚类算法提出了新的挑战，使得聚类算法仍是目前研究的热点问题。

聚类分析时，相似性函数的参数主要来自于对象的属性，常用的是对象间距离函数，如欧氏距离函数（Euclidian Distance Function）、曼哈顿距离函数（Manhattan Distance Function）。

给定 n 个对象，每个对象有 m 个属性，则第 i 个对象可表示为一个 m 维向量：$X_i=(x_{i1}, x_{i2},\cdots, x_{im})$。对象 X_i 和 X_j 之间的欧氏距离函数表示为式（1.1）。

$$D(X_i, X_j)=\frac{1}{m}\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2} \tag{1.1}$$

对象 X_i 和 X_j 之间的曼哈顿距离函数表示为式（1.2）。

$$D(X_i, X_j)=\frac{1}{m}\sum_{k=1}^{m}|x_{ik}-x_{jk}| \tag{1.2}$$

另外，有些应用场合中，也可能使用相关系数或互信息等方法度量对象间的相似性。

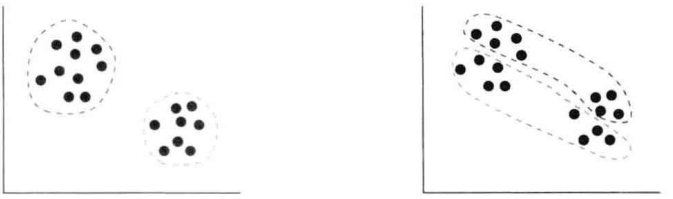
聚类的目标就是根据相似性函数，将对象分到不同的簇中，且使得对象之间的关系满足以下两点：

（1）差异性（Separation）：属于同一个簇的对象高度相似。

（2）同质性（Homogeneity）：分属不同簇的对象相似性低。

同簇内的对象相似性越高，不同簇的对象差异性越大，则聚类质量越好。例如，对于同一组聚类数据，图 1-1（a）所示的聚类质量要好于图 1-1（b）所示的聚类质量。

此为试读，需要完整PDF请访问：www.ertongbook.com



（a）质量好的聚类　　（b）质量差的聚类

图 1-1　聚类质量对比

❑1.2　双向聚类

在聚类分析中，度量对象间的相似性时，若使用了对象的全部属性，此种聚类即是人们通常所指的聚类，为有别于下面的双向聚类，也称为单向聚类。在某些应用中，不宜使用对象的所有属性度量对象间的相似性，使用部分属性度量对象相似性，可能会得到质量更高的聚类结果。例如，基因表达谱表示为矩阵 $A=[a_{ij}]_{m\times n}$，行表示基因，列表示实验条件，矩阵元素 a_{ij} 表示基因 i 在条件 j 下的表达水平，对 A 进行聚类分析的目的是寻找基因簇，同一簇中的基因可能具有相似的功能（详见第3章）。基因表达谱的单向聚类分析无法反映以下两点事实[2-4]。

（1）对于基因表达谱中的基因而言，可能是在部分实验条件下，某些基因的表达是相似的，而不是在所有实验条件下表达是相似的。

（2）一个基因可能会出现在多个簇中，即一个基因可能具有多种生物学功能。

基于以上两点原因，比较对象间的相似性时，可以只使用全部属性的一个子集，而不是使用全部属性。即同时对矩阵的行和列进行聚类，从而避免单向聚类的上述缺点，这种聚类方式称为双向聚类（Biclustering）。

术语双向聚类最早是由 Mirkin 在 1996 年提出的[5]，但这方面最早的研究可见于 Hartigan 等人的文献[6]。双向聚类又称为协同聚类（Co-clustering）、块聚类（Block Clustering）等。在文献[7]中，双向聚类被看作子空间聚类的一种。以下章节中的“聚类”是通称，包含了双向聚类和单向聚类。

2000 年，Cheng 等人首次使用双向聚类方法分析了基因表达谱，实验结

果表明，双向聚类的结果优于单向聚类。其后，分子生物学，尤其是基因聚类分析，成为双向聚类应用最多的领域[7-9]。双向聚类的其他应用领域有：高维数据降维[10]、市场分析[11, 12]、文本挖掘[13]等，更多的应用参见文献[4, 8]。

如前所述，聚类是将拥有某一特征的对象聚到同一个簇中，我们可认为刻画了此特征的对象属性值完整地表达了这个特征，在聚类矩阵中，将与这些属性值对应的行或列恰当地重排后，这些属性值被汇聚在一起，形成一个子矩阵，这个子矩阵称为簇（Cluster）。

粗略地讲，双向聚类的求解目标是在矩阵中寻找多个块状子矩阵，使得在子矩阵所包含的属性下，子矩阵中的对象具有相同特征，这样的子矩阵称为双向簇（Bicluster）。

对于单向聚类分析而言，簇表现为矩阵的条状子矩阵，如图 1-2（a）、（b）所示；对于双向聚类分析而言，双向簇表现为矩阵的块状子矩阵，如图 1-2（c）所示。

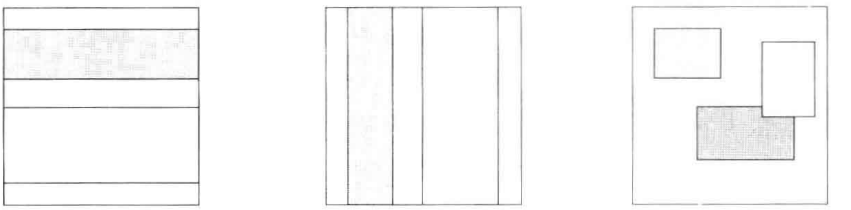


图 1-2　簇与双向簇对应的子矩阵

1.2.1　双向簇的类型

根据双向聚类算法的求解目标，双向簇的类型可分为以下几种。

（1）常数型（如图 1-3（a）所示），簇中元素均为相同的常数。

（2）行常数型（如图 1-3（b）所示），簇中属于同一行的元素是相同的常数。

（3）列常数型（如图 1-3（c）所示），簇中属于同一列的元素是相同的常数。

（4）行数据相关型（如图 1-3（d）所示），簇中的行与行之间呈现出某

种关系，如加关系、乘关系等，图 1-3（d）呈现出的是加关系。

（5）列数据相关型（如图 1-3（e）所示），簇中的列与列之间呈现出某种关系，如加关系、乘关系等，图 1-3（d）呈现出的是加关系。

（6）数据相关演化型（如图 1-3（f）～（j）所示），簇中数据表现为某种顺序（如图 1-3（f）所示），或表现为数据正负变换（如图 1-3（i）所示）等。

上述类型中，第 1～5 种，其相似性评价函数计算的对象是矩阵中元素的数值。第 6 种类型，它的相似性评价函数计算的对象是矩阵中元素的演化行为特征，如在图 1-3（f）～（j）中，矩阵元素是象征性的符号，而不是数据。

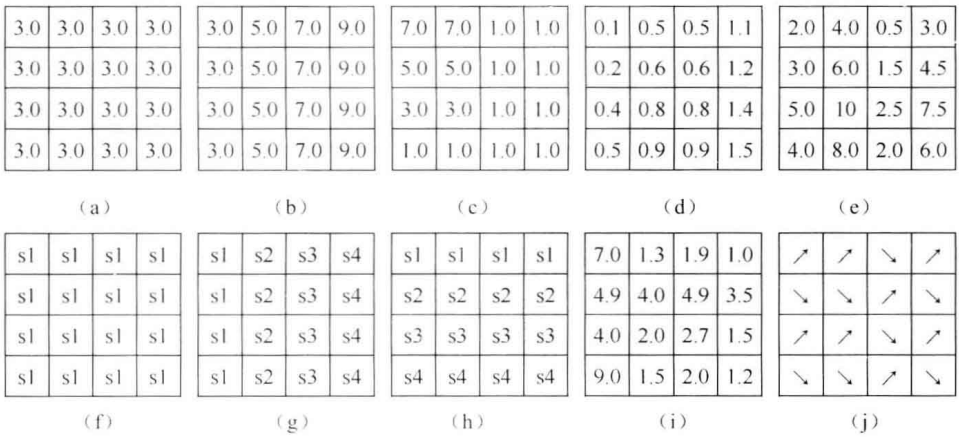


图 1-3　双向簇的类型

1.2.2　双向聚类的解格式

一般地讲，双向聚类算法可以得到多个双向簇，恰当地移动矩阵的行、列后，双向簇在矩阵中是一个块状子矩阵。所谓双向聚类的解格式，就是指这些块状子矩阵在矩阵中的位置关系。早在 1972 年，Hartigan 等人就讨论了双向聚类的解格式问题[6]。Madeira 等人将双向聚类的解格式划分为单一式、行列排他式、行排他式、列排他式、无重叠棋盘式、无重叠非排他式、树结构无重叠式、层级结构重叠式、任意位置重叠式，如图 1-4 所示。本书第 4

章中的双向聚类问题，其解格式是行列排他式。

在行列排他式、行排他式、列排他式、无重叠棋盘式四种解格式中，要求矩阵的行或列至少出现在一个双向簇中，而在单一式、无重叠排他式、树结构无重叠式、层级结构重叠式、任意位置重叠式等解格式中，无此要求。

以计算复杂性的角度研究聚类问题，其计算复杂性是研究的重点内容之一。聚类问题的复杂性与待解具体问题的结构特点密切相关，但无论是双向聚类或是单向聚类，我们感兴趣的聚类问题几乎均是 NP-难的。因此，人们更多的是研究其近似算法，在多项式时间内，计算可接受的问题可行解。

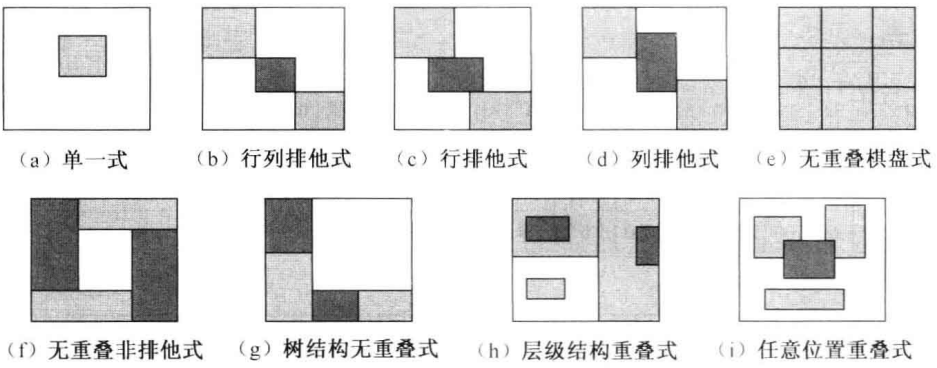


图 1-4　双向聚类算法解格式

❑1.3　数据矩阵上的聚类问题

聚类不存在精确定义[14]。不同文献中的聚类定义依赖于具体应用背景和要解决的问题，例如，自顶向下的聚类认为聚类是一个不断将性质相异的对象逐渐分为多个对象组，且组内对象性质相同的过程[15]；而自底向上的聚类则认为聚类是根据某个自然相似性标准（通常表示为相似性函数），将众多对象分到不同对象分组中的过程，同组内对象相似性高，而不同组内对象相似性低[16]。

在很多应用场合中，聚类问题的实例是数据矩阵，下面给出实数据矩阵下聚类问题的形式化描述。给定 m 个对象和每个对象的 n 个属性值，表示为

矩阵 $A=\left[a_{ij}\right]_{m\times n}$，矩阵中的元素 a_{ij} 表示对象 i 的属性值 j。令 $R=\{1,2,\cdots,m\}$，$C=\{1,2,\cdots,n\}$。

定义 1.1　行向量子集导出子矩阵：令 $R_s\subseteq R$，则称由行向量子集$\{\alpha_i\mid a_{ij}\in\alpha_i,\ i\in R_s,\ j\in C\}$中的元素组成的子矩阵$[a_{ij}\mid i\in R_s, j\in C]$为由行向量子集 R_s 在矩阵 A 中导出的子矩阵，记为 $A'=A[R_s, C]$，也称此子矩阵为行向量子集导出子矩阵。

定义 1.2　列向量子集导出子矩阵：令 $C_t\subseteq C$，则称由列向量子集$\{\beta_j\mid a_{ij}\in\beta_j, i\in R, j\in C_t\}$中的元素组成的子矩阵$[a_{ij}: i\in R, j\in C_t]$为由列向量子集 C_t 在矩阵 A 中导出的子矩阵，记为 $A'=A[R, C_t]$，也称此子矩阵为列向量子集导出子矩阵。

单向聚类的目标是在给定的矩阵中寻找一个以上的行（或列）向量子集导出子矩阵，使得根据相似性评价函数进行比较，子矩阵中的对象（或属性）是相似的，称此子矩阵为簇。

定义 1.3　行列向量子集导出子矩阵：令 $R_s\subseteq R, C_t\subseteq C$，则称由同时属于行向量子集$\{\alpha_i\mid a_{ij}\in\alpha_i, i\in R_s, j\in C\}$和列向量子集$\{\beta_j\mid a_{ij}\in\beta_j, i\in R, j\in C_t\}$的元素组成的子矩阵$[a_{ij}\mid i\in R_s, j\in C_t]$为 R_s 和 C_t 在矩阵 A 中导出的子矩阵，记为 $A'=A[R_s, C_t]$，也称此子矩阵为行列向量子集导出子矩阵。

设 $A_1=A[R_1, C_1]$, $A_2=A[R_2, C_2]$是 A 的两个子矩阵。若 $R_1\cap R_2=\varnothing$，则称 A_1 与 A_2 是行排他的子矩阵；若 $C_1\cap C_2=\varnothing$，则称 A_1 与 A_2 是列排他的子矩阵；若 $R_1\cap R_2=\varnothing$ 且 $C_1\cap C_2=\varnothing$，则称 A_1 与 A_2 是行列排他的子矩阵。

双向聚类的目标是在给定的矩阵中寻找一个以上的行列向量子集导出子矩阵，使得根据相似性评价函数进行比较，子矩阵中的对象（或属性）是相似的，称此子矩阵为双向簇（Bicluster）。

本书第 4 章中的两元矩阵的子矩阵划分问题是一个双向聚类问题。

❑1.4　两元矩阵聚类问题

若矩阵的元素均为 0 或 1，则称此矩阵为两元矩阵。在两元矩阵中，元素均为 1 的子矩阵称为 1 子矩阵，元素均为 0 的子矩阵称为 0 子矩阵。

两元矩阵的聚类分析广泛应用于文本挖掘、市场分析、社区发现等众多

领域。

（1）文本挖掘。给定词项-文档矩阵，行对应于文档，列对应于词，矩阵的元素(i,j)表示词j是否出现在文档i中，1表示出现，0表示未出现。聚类的目标是在此矩阵中寻找1子矩阵。1子矩阵中的文档极有可能属于相同领域。

（2）市场分析。给定顾客-商品矩阵，行对应于顾客，列对应于商品，矩阵的元素(i, j)表示顾客i是否购买了商品j，1表示购买，0表示未购买。聚类的目标是在此矩阵中寻找1子矩阵。1子矩阵中的顾客极有可能有购买相同商品的倾向。

（3）社区发现。给定源对象-目标对象矩阵，行对应于源对象，列对应于目标对象，矩阵的元素(i,j)表示源对象i和目标对象j之间是否存在着某联系，1表示存在，0表示不存在。聚类的目标是在此矩阵中寻找1子矩阵。1子矩阵中源对象与同一组目标对象之间有相同的联系，因此这些源对象极有可能属于同一社区。

另外，在某些应用场合中，将实数矩阵转化为两元矩阵，可以简化计算模型，从而易于对问题复杂性分析，并提高计算速度[3, 17~23]。例如，虽然基因表达谱是一个实数矩阵，但基因表达谱中的基因是否表达，只存在表达或未表达两种可能，故可以将基因表达谱转化为两元矩阵，然后再进行聚类分析，本书第4章中的问题即属于此种情况。在此种情况下，怎样设计恰当的方法两元化矩阵元素是整个问题的关键之一。

来自于真实世界的数据基本上都含有噪声数据，如股票市场上的股票数据、物联网中的传感器数据、基因表达谱等。在两元矩阵中，可以将噪声数据表示为$N \in \{0, 1\}$，这样的两元矩阵称为0-1-N矩阵。对聚类分析0-1-N矩阵时，如何处理噪声数据N是研究重点之一。

❑1.5 割聚类

图聚类算法是一类被广泛研究的重要聚类算法，应用于众多领域[7~9, 24~28]。在图聚类中，聚类对象及其属性表示权重图，聚类的目标是根据问题的要求将图的顶点划分到不同的簇中[24, 25, 29]。大家熟知的图聚类算法包括基于k-

此为试读，需要完整PDF请访问：www.ertongbook.com