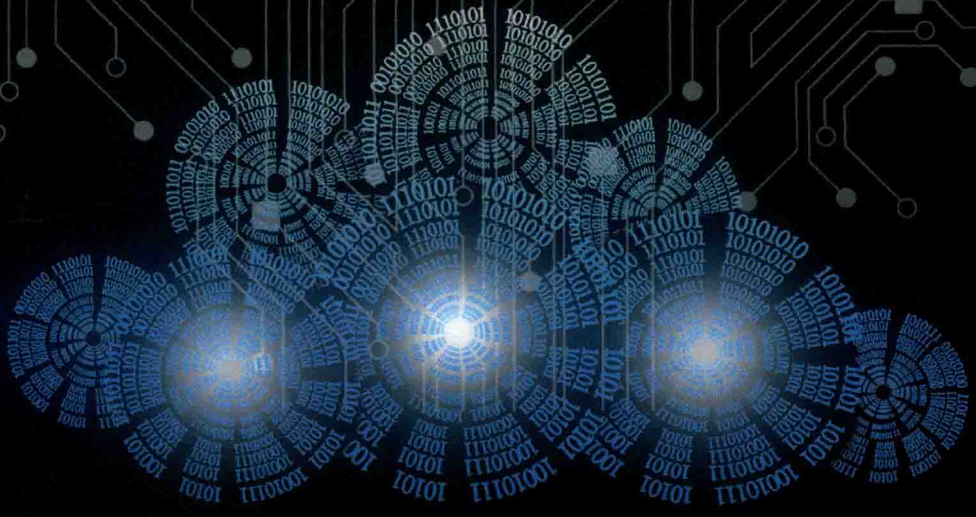
聚类及图聚类流行算法



主　编　陈　梅
副主编　刘春娟　冷明伟

蘭州大學出版社
LANZHOU UNIVERSITY PRESS

国家自然科学基金资助项目（61762057） 资助出版

聚类及图聚类流行算法

主　编　陈　梅
副主编　刘春娟　冷明伟

兰州大学出版社
LANZHOU UNIVERSITY PRESS

图书在版编目（CIP）数据

聚类及图聚类流行算法 / 陈梅主编. -- 兰州 : 兰州大学出版社, 2018.2
ISBN 978-7-311-05330-7

Ⅰ. ①聚… Ⅱ. ①陈… Ⅲ. ①聚类分析－算法－高等学校－教材 Ⅳ. ①O212.4

中国版本图书馆CIP数据核字(2018)第032111号

策划编辑 梁建萍
责任编辑 郝可伟
封面设计 郇 海

书 名 聚类及图聚类流行算法
作 者 陈 梅 主编
出版发行 兰州大学出版社 （地址:兰州市天水南路222号 730000）
电 话 0931-8912613(总编办公室) 0931-8617156(营销中心)
0931-8914298(读者服务部)
网 址 http://press.lzu.edu.cn
电子信箱 press@lzu.edu.cn
印 刷 北京虎彩文化传播有限公司
开 本 787 mm×1092 mm 1/16
印 张 14.5(插页4)
字 数 273千
版 次 2018年3月第1版
印 次 2018年3月第1次印刷
书 号 ISBN 978-7-311-05330-7
定 价 35.00元

作者简介

陈梅，兰州交通大学电子与信息工程学院，博士，副教授。主要从事复杂数据分析及聚类研究，已发表SCI/EI检索文献10余篇，其中以第一作者在世界顶级期刊发表论文2篇。目前主持国家自然科学基金项目1项、甘肃省教育厅高等学校科研项目1项。主持完成甘肃省自然科学基金项目1项、甘肃省财政厅基本科研业务费项目1项。参与完成甘肃省省级、地厅级项目多项。

刘春娟，兰州交通大学电子与信息工程学院副教授，硕士生导师。主要从事信息科学与技术方面的教学与科研工作。主持甘肃省自然科学基金项目、甘肃省建设科技攻关项目、甘肃省高校科研项目3项，参与完成国家自然基金项目1项。发表学术论文10多篇，其中SCI收录2篇，EI收录6篇。

冷明伟，西北民族大学教育科学与技术学院副教授，计算机应用技术专业博士。主要从事数据挖掘和复杂网络方面的研究，已发表聚类和社团检测方向的SCI/EI检索论文20余篇。目前主持国家自然科学基金项目1项、中央高校基金项目1项。主持完成甘肃省教育厅科技课题2项、中央高校基金项目1项、西北民族大学校级课题2项，参与完成甘肃省教育厅科技课题1项。

前 言

我们生活在信息时代，数据的爆炸式增长、广泛可用和巨大数量使得我们的时代成为真正的数据时代。要在这些海量数据中发现有价值的信息、把这些数据转化成有组织的知识，急需功能强大且通用的工具。这种需求导致了数据挖掘的诞生。数据挖掘技术通过对海量数据的挖掘、处理、分析，得出结果，然后给用户提供有价值的“数据”。而“物以类聚，人以群分”，是人类几千年来认识世界和社会的基本方法。如何聚类、分群是从大数据中发现价值必须面对的一个普遍性、基础性问题，是认知科学作为“学科的学科”要解决的首要问题。所以，聚类分析是数据挖掘与大数据分析的关键技术之一。聚类分析指将物理或抽象的数据集合划分为由相似的数据组成的多个簇的分析过程。无论是政治、经济、文学、历史、社会、文化，还是数理、化工、医农、交通、地理、各行各业的大数据或宏观或微观的任何价值发现，无不借助于大数据聚类分析的结果。

但是，除了在数据挖掘类的教材中对聚类技术进行简介外，目前市面上尚未出现系统介绍聚类方法的教材。由此，本书作者在多年系统研究聚类技术的基础上编写了本书。书中从数据的类型、分布、数学基础及常用软件工具开始，首先描述了数据预处理方法、数据可视化方法，然后介绍了经典聚类技术，其中涵盖了基于划分的聚类方法、基于密度的聚类方法、基于层次的聚类方法等，并同时介绍了近几年发表在CCF A类会议和《Science》中几个聚类效果较好的几个新算法，最后对当前研究的热点问题——图聚类算法进行了详细阐述。书中对主要算法皆提供了相关代码、运行方法、运行结果及数据集出处，方便读者学习并进行调试。

本书由陈梅博士担任主编，完成了全书的统筹、编撰工作，并编写了3.2节、3.3节、3.4节、4.7—4.11节。刘春娟老师参与编写了本书的第一章、第二章、3.1节以及4.1—4.6节。冷明伟博士参与了全书的整理与修订工作，并提出了很多建设性的专业的建议和意见。陈梅的硕士研究生林俊山、温晓芳、杨志翀参与了本书的资料收

集、整理以及校对工作。

本书可作为计算机专业、大数据分析与处理专业的本科及研究生教材，也可供从事数据挖掘、大数据分析的科研人员使用。

限于编者水平有限，兼之编书时间仓促，书中难免出现疏漏与欠妥之处，恳请广大读者提出宝贵意见。

编者

2017年11月

目 录

第1章 引论

1.1 认识聚类

随着信息技术的飞速发展，可获得的数据越来越丰富，数据的类型也越来越复杂。衣食住行、各行各业都无不与数据分析紧密相关，一方面，互联网企业需要大量的数据支撑服务体系；另一方面，传统行业需要对过往数据进行分析来提升业绩。对数据的有效分析、利用已成为推动社会发展的重要因素之一。然而，很多时候，海量的、复杂的数据让人眼花缭乱，无从下手，给人们的认知造成了很大的困扰。很多企业甚至不能对收集到的庞大的数据信息进行很好的处理和分析。“得数据者得天下”，但是直接把海量数据推给用户是毫无意义的。这就需要通过对海量数据进行挖掘、处理、分析，得出结果，找出隐藏在数据中的可用信息，从而为用户提供有价值的数据。

这时候，聚类技术作为根据对象的特征将对象集合分成由类似的对象组成的多个类的分析过程，就提供了一个很好的选择。聚类分析是一类重要的人类行为，早在孩提时代，一个人就能通过不断改进下意识中的聚类模式来学习如何区分动物、植物。根据事物的特征对其进行聚类或分类，可以从大量数据中提取隐含的、未知的、有潜在应用价值的信息或模式。“物以类聚，人以群分”，这是人类几千年来认识世界和社会的基本方法。如何聚类是从大量数据中发现其价值必须面对的一个普遍性、基础性问题，是认知科学作为“学科的学科”要解决的首要问题。“无论是政治、经济、文学、历史、社会、文化，还是数理、化工、医农、交通、地理、各行各业的大数据或宏观或微观的任何价值发现，无不借助于大数据聚类分析的结果，因此，数据分析和挖掘的首要问题是聚类，这种聚类是跨学科、跨领域、跨媒体的。如何进行大数据聚类是数据密集型科学的基础性、普遍性问题。”所以，聚类将会成为数据认知的突破口。聚类是挖掘数据资产价值的重要一步，可以让我们主动迎接信息化时代，直面信息化带来的挑战。

实际生活中，我们可以借助聚类对数据做出深层次的挖掘，做出归纳性的推理，从中挖掘出潜在的模式，认识海量数据可能带来的深刻影响和巨大价值，改变我们的生活、工作和思维方式。在商务领域，聚类分析能帮助市场分析人员从客户数据库中发现不同的客户群，用购买模式来刻画不同客户群的特征。利用聚类算法从大量数据中挖掘出的有用模式，将会应用于我们的生活，为我们的生活提供便利。比如，在健康方面，我们可以利用智能手环监测的数据，对睡眠数据进行聚类，进而分析睡眠模式，了解我们的睡眠质量；在汽车保险方面，如果采集了汽车的每一次行驶信息、每一次维修信息、每一次刹车信息，通过对这些数据进行聚类分析，保险公司可对一个车况好、驾驶习惯好、常走线路事故率低、不勤开车的特定客户，给予更大的优惠，而对风险太高的客户提高保险报价甚至拒绝其投保；在保障房购买上，可采集保障房申请人群的收入、工作、身体状况以及年龄信息，通过聚类分析帮助发现最需要保障房的人群；在生物学上，聚类能用于分析植物和动物的基因信息，获得对种群中固有结构的认识。

总之，在信息化时代，研究聚类显得尤为重要。聚类作为一门蓬勃发展的技术，将会成为数据认知的突破口，成为很多行业的核心竞争力。本书的研究工作致力于提出兼顾效率和有效性的聚类算法，使聚类在确保精确性的同时，在海量数据背景下也能以较低的时间复杂度高效运行。

1.2 聚类分析概述

数据挖掘是从大量的数据中通过算法发现隐含在其中的有价值的、潜在有用的信息和知识的过程，也是一种决策支持过程，其主要基于人工智能、机器学习、模式识别以及统计学等。数据挖掘最常用的方法有分类、聚类、预测、回归分析、关联规则等。聚类是多元数据分析的主要方法之一，是数据挖掘采用的一项关键技术。

聚类分析起源于分类学，但是聚类与分类又有明显的不同。分类是在已知类别标号的情况下，将其他数据点映射到给定类别的某一类。然而，在很多情况下，数据的类别标号是未知的，却又需要对其进行分组。这时候，就需要借助于聚类。聚类分析能够根据数据相似度自动发现数据的分组、挖掘出数据中潜在的数据模式、特征以及规律。因此，在机器学习领域的研究中，聚类被认为是一种无监督的学习过程。

聚类是根据数据间的相似性把一个数据集划分成多个组或簇的过程，使得同一簇内的数据尽可能相似，而与其他簇内的数据尽可能不相似，也就是说，让同一簇

内的数据分布尽可能紧凑，而不同簇间的数据尽可能远离。相似性一般根据对象的属性值进行评估，紧凑性根据数据间的距离来衡量。两个数据间的相似度值越高，它们之间就越相似。而距离则正好相反，两个数据间的距离越远，它们越不相似。因此，距离度量也被称为相异性度量。

为了适应不同特征数据的应用需求，近几十年来，研究者提出了大量的基于不同理论的聚类算法。一般而言，聚类算法可以划分为以下几类[1]：基于划分的方法、基于密度的方法、基于层次的方法、基于网格的方法和基于模式的方法。很多算法中，这些类别可能相互重叠，一种算法可能同时具有几种方法的特征。

作为数据挖掘的一种强有力的分析工具，聚类分析一般具有两种用途：

（1）作为一种独立的数据挖掘工具，发现数据的分布特征；

（2）作为其他一些数据分析方法的数据预处理步骤，给其他方法提供基于某种模式已进行了分组的数据，进一步让其他方法在相应的数据划分结果上进行专业的分析。

目前，聚类分析已经成功应用于许多领域，包括图像处理、模式识别、商业、生物、地理、网络服务、情报检索等。通过对数据进行聚类分析，可以把隐没于一大批看似杂乱无章的数据中的信息集中、萃取和提炼出来，以找出所研究对象的内在规律，从中挖掘出潜在的模式，还可以帮助企业、商家调整市场政策、减少风险、理性面对市场，并做出正确的决策，也可以帮助政府调整未来的管理政策、经济结构，积极应对生态发展等。

截至目前，研究人员已经提出了不同种类、面向各种数据特征的聚类算法。为了对众多算法进行比较分析，从而选择出最合适的聚类分析算法，人们从外部评价和内部评价两个方面提出了一些聚类评价方法。其中外部评价是依据标准的数据划分对聚类算法的结果进行质量评价；内部评价是根据簇内数据自身的分布对算法的聚类结果进行评价。

1.3 参考文献

[1] 韩家炜，坎伯，裴健.数据挖掘：概念与技术［M］.北京：机械工业出版社，2012.

第2章 聚类基础

2.1 聚类定义

聚类分析的目的就是在未知的情况下将相似的数据划分到一起，不相似的数据分开。聚类分析跨越多个领域，包括数学、计算机科学、统计学、生物学和经济学等。在不同的应用领域，很多聚类技术都得到了发展，这些技术方法被用于描述数据、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[1]。

从统计学的角度看，聚类分析是通过数据建模简化数据的一种方法。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类等。采用k-Means、k-Medoids等算法的聚类分析工具已被加入许多著名的统计分析软件包中，如SPSS、SAS等。

从机器学习的观点讲，簇相当于隐藏模式。聚类是搜索簇的无监督学习过程。与分类不同，无监督学习不依赖预先定义的类或带类标记的训练实例，需要由聚类学习算法自动确定标记，而分类学习的实例或数据对象有类别标记。聚类是观察式学习，而不是示例式学习。

聚类分析是一种探索性的分析，在聚类的过程中，人们不必事先给出一个分类的标准，聚类分析根据数据间的相关关系，自动进行分类。不同的聚类分析方法，得到的结论经常会不同。不同的研究者对同一个数据集进行聚类分析，所得到的聚类数可能不尽相同。

就实际应用来看，聚类分析有三个作用，分别为：

（1）聚类分析属于数据挖掘的主要任务之一。

（2）聚类分析可以作为其他算法（如分类和定性归纳算法）的预处理步骤。

（3）聚类分析能够作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的簇集做进一步分析。

2.1.1 形式化定义

聚类是一种把数据集划分成簇的过程，并使得簇内的点尽可能相似，而簇间的点差异性尽可能大。也就是说，同一类的数据点尽可能被聚集到一起，同时让不同类的数据点尽量分离。挖掘出的每一个簇都可能包含着某种潜在的数据模式、特征以及规律。

聚类是将一个数据集划分成子集的过程。在本书中，数据集D被定义为，

$$D=\{x_1,x_2,\cdots,x_i,\cdots x_n\} \tag{2.1}$$

其中，$x_i(1<i<n)$是数据集中的第i个数据点，n是数据集D内数据点的个数。D也可以用如下的$n\times p$矩阵表示，

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & & & & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & & & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

其中，p是每个数据点的维数，矩阵的每一行对应一个数据点x_i，而x_{if}表示D的第i个数据点x_i的第f个属性。

本章中，聚类用如下方式进行定义：

聚类算法将数据集D划分成$k(1\leqslant k\leqslant n)$个簇，$C_1,C_2,\cdots,C_k$

（1）$C_i\neq\varnothing,i=1,\cdots,k$

（2）$C_i\subseteq D,i=1,\cdots,k$

（3）$\bigcup_{i=1}^{k}C_i=D$

（4）$C_i\cap C_j=\varnothing,1\leqslant i,j\leqslant k,i\neq j$

其中，每个C_i内的点相互相似，而与C_j内的点不相似。

2.1.2 相似度计算

数据间的相似度是聚类分析中一个非常重要的概念。无论是在聚类过程中还是在最终的聚类质量评价中，相似度都起着至关重要的作用。最常用的两点间相似性的度量方式是距离度量和相似性度量。其中，距离度量以数据集D中点x_i与x_j之间的距离$d(x_i,x_j)$来度量两点间的相似程度，$d(x_i,x_j)$越大，x_i与x_j越不相似，而$d(x_i,x_j)$越小，x_i与x_j就越相似。相似性度量直接以两点间的相似程度$sim(x_i,x_j)$作为度量的基

础，$sim(x_i,x_j)$越大，x_i与x_j越相似，反之亦然。本节将介绍几个常用的距离函数和相似性函数。

1.常用的距离函数

（1）欧氏距离。欧氏距离是最常用的相似度度量距离，很多聚类算法预设的距离都是欧氏距离。它常用来表示两点间的最短距离，即直线距离：

$$d(x_i,x_j)=\sqrt{\sum_{f=1}^{p}(x_{if}-x_{jf})^2} \tag{2.2}$$

（2）曼哈顿距离。曼哈顿距离也叫城市街区距离。它表示两点的所有属性间的距离的总和。通俗地讲，它表示的是城市两点之间的街区距离：

$$d(x_i,x_j)=\sum_{f=1}^{p}|x_{if}-x_{jf}| \tag{2.3}$$

（3）闵科夫斯基距离。闵科夫斯基距离是欧氏距离和曼哈顿距离的推广：

$$d(x_i,x_j)=\sqrt[h]{\sum_{f=1}^{p}|x_{if}-x_{jf}|^h} \tag{2.4}$$

其中，$h\geqslant 1$。

可以看出，当h=1时，它是曼哈顿距离；当h=2时，它是欧氏距离。

（4）切比雪夫距离。切比雪夫距离又称为上确界距离，常被记为$L_{\max}$或者L_∞。它取两点间所有维数距离的最大值：

$$d(x_i,x_j)=\max_f|x_{if}-x_{jf}| \tag{2.5}$$

2.常用的相似度函数

（1）余弦相似性。余弦相似性度量两个向量$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间夹角的余弦：

$$sim(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}_j}{||\boldsymbol{x}_i||\cdot||\boldsymbol{x}_j||} \tag{2.6}$$

其中$||\boldsymbol{x}_i||$是向量$\boldsymbol{x}_i$的长度，定义为$\sqrt{x_{i1}^2+x_{i2}^2+\cdots+x_{ip}^2}$。

（2）皮尔逊相关系数。皮尔逊相关系数度量首先将两个数据点x_i和x_j做Z-score处理，然后将两组数据的乘积除以数据的维数：

$$sim(x_i,x_j)=\frac{Z(x_i)\cdot Z(x_j)}{p} \tag{2.7}$$

（3）Jaccard相关系数。两个数据点x_i和x_j的Jaccard相关系数定义如下：

$$sim(x_i,x_j)=\frac{x_i^{\mathrm{T}}\cdot x_j}{\|x_i\|^2+\|x_j\|^2+\|x_i\|\cdot\|x_j\|} \tag{2.8}$$

可以看出，除了切比雪夫距离，以上所有的距离和相似性函数都与数据的维数紧密相关。

2.1.3 聚类结果评价

实际应用中，我们经常需要判断聚类算法产生的聚类结果是否合理，并需要比较多个算法在同一个数据集上产生的聚类结果的优劣。这就需要对聚类的质量进行评价。另外，对聚类结果进行评价还可以帮助估计算法在每个数据集上的聚类可行性和实用性，帮助确定聚类算法的输入参数。聚类评价分为外在评价和内在评价[2,3]两种方法。

1. 外在评价方法

在已知数据集标准划分的情况下，利用标准划分评价某种算法对数据划分的结果，就称为外部评价方法。最常用的外部评价方法有 *Rand Index*（RI）[4]、*Adjusted Rand Index*（ARI）[5]、*Normalized Mutual Information*（NMI）[6]、*F-measure*（亦即 *F*-score）[7]、*Jaccard Index* 和 *Purity*。

为了方便以下描述，本书将数据集的标准划分记为$S=\{S_1,S_2,\cdots,S_s\}$，待评价的某种算法的划分记为$P=\{P_1,P_2,\cdots,P_m\}$。

（1）*Rand Index*

用a,b,c,d表示两个点x_i和x_j相应的簇分配：

$$a=\left|\left\{(x_i,x_j)|x_i,x_j\in S_k,x_i,x_j\in P_l\right\}\right|$$

$$b=\left|\left\{(x_i,x_j)|x_i\in S_{k1},x_j\in S_{k2},x_i\in P_{l1},x_j\in P_{l2}\right\}\right|$$

$$c=\left|\left\{(x_i,x_j)|x_i,x_j\in S_k,x_i\in P_{l1},x_j\in P_{l2}\right\}\right|$$

$$d=\left|\left\{(x_i,x_j)|x_i\in S_{k1},x_j\in S_{k2},x_i,x_j\in P_l\right\}\right|$$

其中，$1\leqslant i,j\leqslant n;i\neq j;1\leqslant k,k_1,k_2\leqslant s;k_1\neq k_2;1\leqslant l;l_1,l_2\leqslant m;l_1\neq l_2$。

那么，*Rand Index* 可表示如下：

$$RI=\frac{a+b}{a+b+c+d}=\frac{a+b}{\binom{n}{2}} \tag{2.9}$$

直觉$a+b$可被看作是S和P两个划分的一致性，$c+d$可被看作是S和P两个划分的

偏差。*Rand Index*的取值范围为[0,1]，0意味着两个划分完全不同，1则表示两个划分完全一致。其值越大，类的划分就越好。

（2）*Adjusted Rand Index*

*Adjusted Rand Index*是*Rand Index*的调整形式。

$$ARI = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} \tag{2.10}$$

为了更具体地描述*ARI*，用表2.1表示基准划分S和实际划分P一致的部分，其中n_{ij}是既在S_i又在P_j中点的数目，亦即$n_{ij} = S_i \cap P_j$。

表2.1 n_{ij}构成的列联表

$S \setminus P$	P_1	P_2	...	P_m	*Sums*
S_1	n_{11}	n_{12}	...	n_{1m}	a_1
S_2	n_{21}	n_{22}	...	n_{2m}	a_2
...	...	...	...	...	...
S_s	n_{s1}	n_{s2}	...	n_{sm}	a_s
Sums	b_1	b_2	...	b_m	

则*Adjusted Rand Index*可表示为

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{2.11}$$

其中n_{ij}，a_i，b_j的值可由表2.1得到。

从公式（2.10）可以看出，当*Index*的值比*ExpectedIndex*的值小时，*ARI*可能为负值。*ARI*的取值范围为［-1，1］，*ARI*值越大，类的划分越好。

（3）*Normalized Mutual Information*

*Normalized Mutual Information*是一个基于信息论的衡量标准：

$$NMI = \frac{I(S,P)}{\frac{H(S) + H(P)}{2}} \tag{2.12}$$

其中，$I(S,P)$为标准划分S与实际划分P的互信息量，可用来评价两种划分结果的一致性；$H(S)$和$H(P)$分别表示这两种划分的熵。

更进一步，*NMI*可用概率表示为：