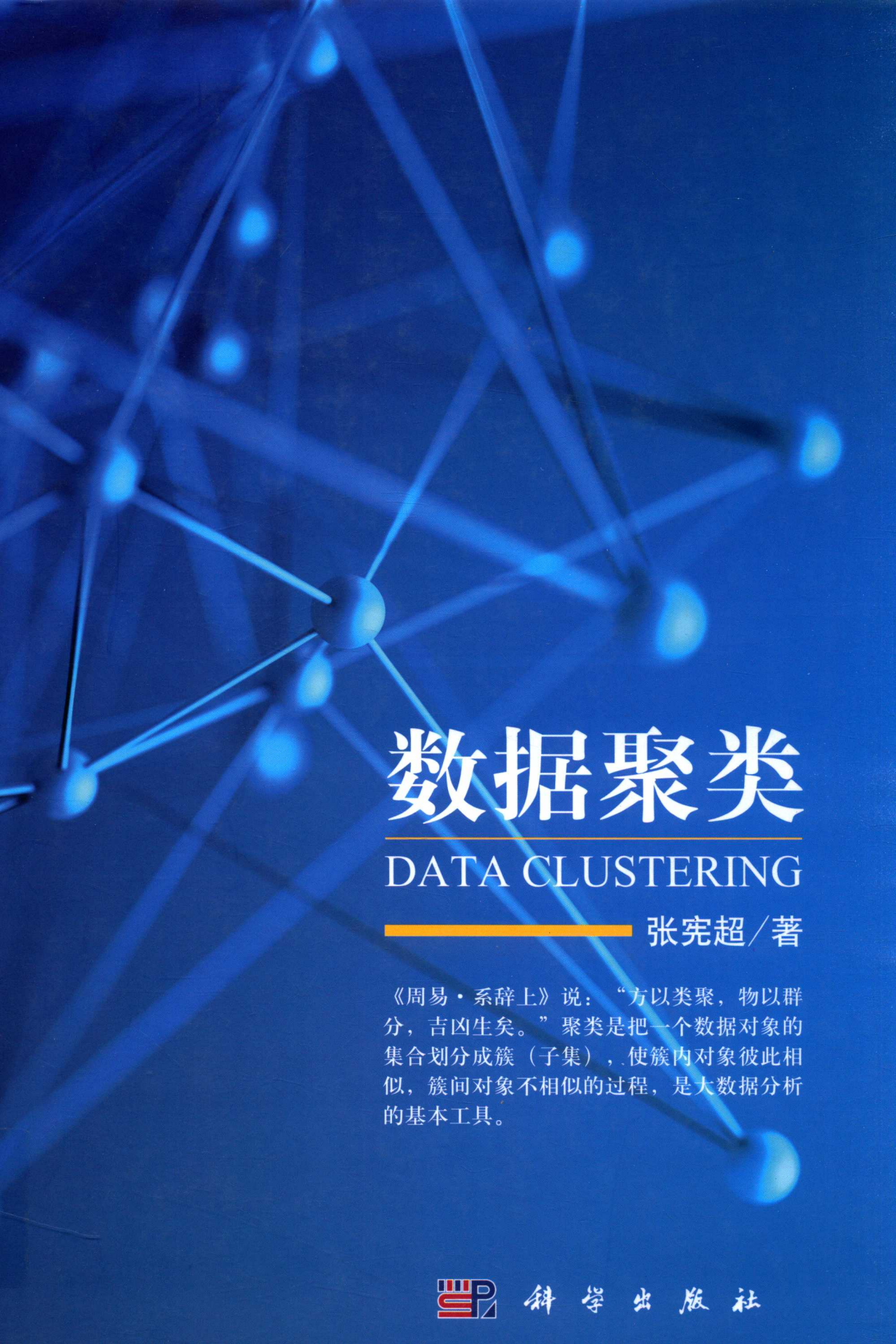

数据聚类

DATA CLUSTERING

张宪超/著

《周易·系辞上》说："方以类聚，物以群分，吉凶生矣。"聚类是把一个数据对象的集合划分成簇（子集），使簇内对象彼此相似，簇间对象不相似的过程，是大数据分析的基本工具。

科学出版社

数据聚类

张宪超 著

科学出版社
北京

内 容 简 介

聚类是数据挖掘领域的一个重要分支。本书全面系统地介绍聚类的主要方法。首先，对涉及聚类的各个方面进行简略的综述；然后，对各类聚类算法进行较详细的讨论。本书主要内容分为三大部分：第一部分是经典算法部分（第 2～6 章），讨论 k-均值、DBSCAN 等传统算法；第二部分是高级算法部分（第 7～12 章），讨论半监督聚类、高维数据聚类、不确定数据聚类等；第三部分是多源数据聚类部分（第 13 章），主要讨论多视角聚类和多任务聚类。

本书可供数据科学与人工智能等领域的研究人员、工程技术人员、相关学科研究生和基础较好的高年级本科生参考阅读。

图书在版编目(CIP)数据

数据聚类/张宪超著. —北京：科学出版社，2017.5

ISBN 978-7-03-052846-9

Ⅰ. ①数… Ⅱ. ①张… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 110688 号

责任编辑：张 震 杨慎欣 / 责任校对：彭 涛

责任印制：张 倩 / 封面设计：无极书装

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

中国科学院印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2017 年 5 月第 一 版 开本：787×1092 1/16

2017 年 5 月第一次印刷 印张：25 1/2 插页：3

字数：666 000

定价：188.00 元

（如有印装质量问题，我社负责调换）

序

2007年，关系型数据库鼻祖、图灵奖得主Jim Gray提出了科学研究的第四范式，即数据密集型科学发现。在此之前，人类认知世界的方式经历了三个阶段的飞跃：第一范式是实验科学，从原始的钻木取火发展到后来以伽利略为代表的文艺复兴时期的现代科学发展初级阶段；第二范式是理论科学，从经典的牛顿力学到量子力学和相对论，人们以超凡的头脑思考超越了实验设计；第三范式是计算科学，20世纪中叶以来，人们通过对复杂现象进行计算机模拟仿真实验，推演出越来越复杂的现象，如模拟核试验、天气预报等。

第四范式是指随着数据的爆炸性增长，计算机将不仅仅能做仿真实验，还能进行分析总结。科学的第四范式必将推动人类认知世界能力的一次新的革命。近年来大数据概念的火热正是科学第四范式开始发挥作用的集中反映。大数据不仅引起学术界和产业界的普遍重视，更上升为世界各国的国家战略。然而，要充分发挥大数据的作用，必须具备强大的数据分析能力。因此，与大数据概念火爆同步的是人工智能的再度爆发。大数据和人工智能是科学第四范式必不可少的两个方面：没有人工智能，大数据是埋在深山里的金矿；没有大数据，人工智能是巧妇难为无米之炊。

人工智能的核心是机器学习，其首要任务是对事物进行辨识和区分。机器学习分为监督学习和无监督学习两大类。监督学习的主要任务是分类，即用大量已标记数据完成对新数据的区分；无监督学习的主要任务是聚类，即在没有任何人工干预情况下对数据进行区分。值得高兴的是，近几年来，随着人们对传统神经网络技术的深度挖掘，即深度学习的发展，使监督学习的能力得到了飞跃式的进步。但是，在这里我还是要强调无监督学习的作用：正如2015年深度学习的领军人物Lecun、Bengio和Hinton在*Nature*上的综述所说，“人和动物的学习很大程度上是无监督的：我们通过观察发现世界的结构，而不是对每个物体命名”，无监督学习应该是大数据分析最基本的工具。人们必须清楚地看到，无监督学习的难度远大于监督学习，无监督学习的研究者远少于监督学习的研究者，因此，无监督学习的进展相对缓慢。用一个龟兔赛跑的例子来比喻：监督学习的发展像兔子跳了几次，无监督学习的发展则像乌龟缓慢而稳步地前行。

尽管如此，几十年来仍有学者在无监督学习领域孜孜不倦地进行探索，研究出包括k-均值算法在内的很多著名成果。尤其近年来，随着人们对无监督学习重要性的认识，也有更多的学者投入其中，取得了很多突破性进展，包括高维数据聚类、不确定数据聚类、多源数据聚类等。及时对前期研究进行梳理和总结是十分必要的，《数据聚类》因此应运而生。

《数据聚类》的作者曾是我的学生，我对他颇为熟悉。他是一个敢于挑战难题的青年学者。他的博士论文曾以最大流算法为课题，而今，他又从研究最大流最小割理论过渡到图聚类进而一般聚类问题的研究。他在聚类领域已经坚持了十余年，在这个年代如此投入地做基础研究十分难能可贵。而作者的成绩也是斐然的，他在高维数据聚类、谱聚类、不确定数据聚类

等方面都有很好的成果，在多视角聚类和多任务聚类方面更是取得了可喜的成果，尤其是在国际上率先提出了多任务多视角聚类问题和算法。《数据聚类》从算法角度对聚类问题进行了全面的梳理，视角独特而论述新颖，包含很多作者自己的研究工作，是一本很有价值的著作。

中国科学院院士
首届国家级教学名师　陈国良

2016 年 11 月

前言

2016 年初，谷歌围棋 AlphaGo 以 4∶1 的成绩战胜了人类围棋世界冠军李世石，引起全世界的关注，这标志着人工智能的发展进入了一个全新的阶段。近几年来，人工智能得到飞速的发展，在很多领域如图像识别、语音识别等方面取得了突破性的进步。人工智能的研究也得到全世界学术界和产业界的高度关注，进入了一个新的高潮期。种种迹象表明，人类进入全方位智能时代已经为期不远了。所有这一切几乎均得益于神经网络的新技术——深度学习的发现和发展（非常有趣的是人工智能的几次高潮均来自神经网络的进步，可见神经网络的生命力）。深度学习的概念由 Hinton 等于 2006 年提出，在近年来已经逐渐成为机器学习的主流技术，在多数应用领域的性能明显超出已有技术。

机器学习包括监督学习和无监督学习。目前的深度学习基本上只带来监督学习的进步，但仅靠监督学习是无法实现完整的人工智能的。作为智能系统，监督学习似乎足够“能”而不足够“智”。足够“能”体现为它能够在大数据中挖掘知识，这甚至是人脑做不到的。事实上人脑并不是处理大数据的系统，人类在任何领域所掌握的知识均有限，例如，每个人仅认识数千个汉字或单词。不足够“智”体现为监督学习需要大量人工标记的训练样本。人脑的学习并不需要大量的样本训练，人类是在没有指导或少量指导的条件下获得知识的，而且人脑会不断地学习并强化自己在各个领域的知识。人类在有限知识的基础上体现出惊人的创造力。类似人脑的智能系统更需要无监督学习、小样本学习、强化学习和迁移学习等功能。因此，人工智能的发展仍然任重而道远。

本书讨论聚类技术。聚类是无监督学习的主要内容，在很多文献中人们甚至把聚类和无监督学习两个概念等价使用。聚类一直是机器学习、数据挖掘、模式识别等领域的重要组成内容，近年来更得到高度重视。2015 年，中国人工智能学会理事长李德毅院士在“新一代信息技术产业发展高峰论坛”上指出：“人类的认知科学要想有所突破，首先就要在大数据聚类上取得突破，聚类是挖掘大数据资产价值的第一步。”同年，深度学习的领军人物 Lecun、Bengio 和 Hinton 在 *Nature* 上的综述指出：“人和动物的学习很大程度上是无监督的：我们通过观察发现世界的结构，而不是对每个物体命名。”

那么什么是聚类呢？《周易 • 系辞上》说：“方以类聚，物以群分，吉凶生矣。”自然的事物总是按一定的规律组织起来的，人们通过认识这些组织的结构特征获得知识，从而做出决策。以生物为例（我们这个世界是因为有生物而活泼生动的），人们根据生物的相似程度（包括形态结构和生理功能等），把生物划分为种和属等不同的等级，并对每一类群的形态结构和生理功能等特征进行科学的描述，以弄清不同类群之间的亲缘关系和进化关系。相信很多人小时候学习生物时都会惊讶于鲸居然是哺乳动物而不是鱼，猫和老虎是同一科等。

和分类（监督学习的主要任务）不同，聚类是在无标记样本的条件下将数据分组，从而发现数据的天然结构。聚类在数据分析中扮演重要的角色，它通常被用于以下三个方面。

（1）发现数据的潜在结构：深入洞察数据、产生假设、检测异常、确定主要特征。

（2）对数据进行自然分组：确定不同组织之间的相似程度（系统关系）。

（3）对数据进行压缩：将聚类原型作为组织和概括数据的方法。

这几个方面的功能使聚类既可以作为预处理程序，又可以作为独立的数据分析工具。

聚类是典型的交叉学科，在很多领域有广泛的应用，其研究已有 60 多年的历史。生物分类学者、社会学者、哲学家、生物学家、统计学家、数学家、工程师、计算机科学家、医学研究者等众多收集和处理实际数据的工作者都对聚类方法做出了贡献。在不同的领域，聚类还可能被称为 Q-分析、拓扑、凝结、分类等。聚类的概念最早出现在 1954 年的一篇处理人类学数据的论文中。自此开始，聚类一直是相关领域重要的研究内容之一。2009 年，有人用谷歌学术搜索做过统计，发现仅 2007 年一年就有 1660 个包含“数据聚类”的条目。几十年来有数以万计的文献讨论聚类算法及其在科学和工程领域的应用，这充分说明聚类对数据分析的重要性。

人们已经发表了数千种聚类算法。总体来说，这些算法分为基于划分的、基于层次的、基于密度的等几大类。目前，最流行并且简单的算法是于 1955 年发表的 k-均值。尽管 k-均值已经发布了 60 多年，期间有数千种其他算法出现，它仍然是应用最广的算法。这也说明了设计一般聚类算法的难度和聚类问题的不适定性。近年来，高维数据、不确定数据、多模数据的大量出现又给聚类带来很多新的挑战。同时，大数据的一个显著特征是多源性，能否从多源互补的数据中获得有价值的信息又是全新的课题。

总之，人们已经就聚类做了大量的工作，聚类问题的研究取得了显著的进步。不过和神经网络技术的几起几落不同的是，聚类的研究是稳步而缓慢的，至今仍有很多问题亟待解决。适时对前期工作进行较全面的总结，补充最新的研究内容，并且对进一步工作进行梳理，是十分必要的。

关于聚类的研究可以大体分为三个方面。①以技术为中心的研究：聚类是个范围很广的课题，有大量工作讨论如何选择特征、如何进行测度学习、如何对数据进行划分等。②以数据为中心的研究：不同的应用领域产生不同的数据类型，而不同的数据类型会导致不同的测度和启发规则的选择。③聚类衍生问题研究：包括数据可聚类性、聚类特征选择、聚类可视化、聚类验证、集成聚类等。

本书旨在从技术角度对聚类问题的研究进行梳理和讨论。首先对涉及聚类的各个方面进行简略综述，接下来对各类聚类算法进行较详细的讨论。本书主要内容可以分为三大部分：第一部分是经典算法部分，讨论 k-均值、DBSCAN 等传统算法，这些算法大多出自 2000 年以前，但本书会对这些算法的最新进展进行补充；第二部分是高级算法部分，讨论高维数据聚类、不确定数据聚类、协同聚类、半监督聚类等，这些算法主要出现在 2000 年以后；第三部分是多源数据聚类部分，主要讨论多视角聚类和多任务聚类，这些问题在最近几年成为热点研究课题。

作为机器学习和数据挖掘的重要组成部分，聚类和其他学习算法一样建立在较为复杂的数学工具基础之上，这要求读者具备一定的数学基础，尤其是线性代数、概率论与数理统计、数值计算和优化方法等。同时要求读者具备数据库、数据结构和算法的基本知识和某种程序设计语言（如 C、MATLAB 等）的基本运用能力。

聚类是个很大的课题，从海量的研究中梳理和总结不是一件容易的事。笔者本着自己的理解，尽最大努力为读者提供较全面和系统的参考。然而限于水平，难免有疏忽和遗漏，很多观点也仅是笔者的一家之言，不当之处还望读者不吝赐教。深表谢意。

张宪超

2016 年 11 月于大连燕南园

符 号 表

$\boldsymbol{R}^n$	n 维实数向量空间
H	希尔伯特空间
D	数据样本（数据集）
n	数据样本的个数
d	数据样本的维度
k	类的个数
x	标量
$\boldsymbol{x}$	向量
$\boldsymbol{X}$	矩阵
$\boldsymbol{I}$	单位矩阵
$\boldsymbol{X}^{\mathrm{T}}$	转置
$\boldsymbol{X}^{-1}$	逆矩阵
$\|\boldsymbol{X}\|_p$	L_p 范数， p 缺省时为 L_2 范数
X	随机变量
$\mathcal{D}$	概率分布
$P(X)$	概率质量函数， X 是离散的
$p(X)$	概率密度函数， X 是连续的
$P(X,Y)$	联合概率分布
$p(X \mid Y)$	条件概率密度函数
$E[X]$	随机变量的期望值
$\mathrm{Var}(X)$	方差
$\mathrm{Cov}(X,Y)$	协方差
$l(\theta \mid \mathcal{X})$	样本 $\mathcal{X}$ 上具有参数 θ 的似然函数
$\mathcal{L}(\theta \mid \mathcal{X})$	样本 $\mathcal{X}$ 上具有参数 θ 的对数似然函数
$\{\cdots\}$	集合
$\lvert\{\cdots\}\rvert$	集合中元素个数

$\{x_i\}_{i=1}^{N}$	x 的集合，下标 i 遍取1到 N
$\mathrm{I}(x)$	如果 x 为真，则值为1；否则为0
$\mathrm{sgn}(x)$	如果 $x<0$，则值为 -1；如果 $x>0$，则值为1；否则为0
δ_{ij}	如果 $i=j$，则值为1；否则为0
$g(x\mid\theta)$	x 的函数，其定义依赖于参数 θ
$\arg\max_{\theta} g(x\mid\theta)$	函数 g 关于参数 θ 的最大值
$\arg\min_{\theta} g(x\mid\theta)$	函数 g 关于参数 θ 的最小值
$K(x,y)$	核函数

目　录