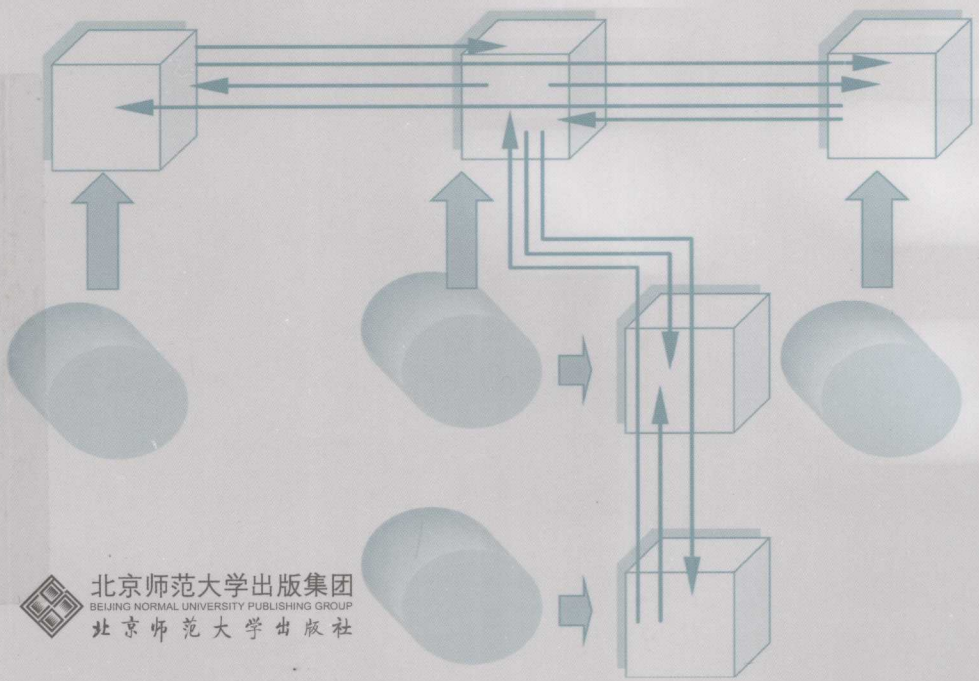
国家自然科学基金资助（60775032）
北京师范大学校级重点学科资助

基于知识的聚类

从数据到信息粒

Knowledge-Based Clustering
From Data to Information Granules

[加]Witold Pedrycz 著
于福生 译



北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

基于知识的聚类

从数据到信息粒

Knowledge-Based Clustering

From Data to Information Granules

[加]Witold Pedrycz 著

于福生 译

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目(CIP)数据

基于知识的聚类　从数据到信息粒/(加)派垂驰(Pedrycz, W.)著，于福生译. —北京：北京师范大学出版社，2008.12
ISBN 978-7-303-09692-3

Ⅰ.基… Ⅱ.①派…②于… Ⅲ.软计算粒计算—模糊系统研究 Ⅳ.N94

中国版本图书馆 CIP 数据核字(2008)第 175331 号
北京市版权局著作权合同登记图字：01-2007-1119

出版发行：北京师范大学出版社 www.bnup.com.cn
北京新街口外大街 19 号
邮政编码：100875
印　　刷：北京新丰印刷厂
经　　销：全国新华书店
开　　本：170 mm × 230 mm
印　　张：19.25
字　　数：340 千字
印　　数：1～2 000 册
版　　次：2008 年 12 月第 1 版
印　　次：2008 年 12 月第 1 次印刷
定　　价：39.00 元

责任编辑：岳昌庆　　装帧设计：高　霞
责任校对：李　菡　　责任印制：李　丽

中译本作者的话

粒计算和词语计算是模拟人类粒化信息、粒度使用信息行为的理论，它将人们对数据的关注从点的层面提升到集合的层面. 在这些理论中，作为基本的概念和运算对象，信息粒起着关键的作用. 作为一种重要的信息粒建立方法和数据分析方法，聚类分析受到人们极大关注，模糊聚类因允许不同程度的类隶属更是得到了长足发展和广泛应用. 不同于以数据为中心的聚类分析方法，W. Pedrycz 所著的《基于知识的聚类：从数据到信息粒》一书的核心内容——基于知识的聚类，论述的是一类以人为中心的模糊聚类分析方法，它强调了人的知识在聚类中的指导作用. 这是一本与众不同、特色鲜明的书，书中所呈现的方法既具有典型性，又具有启发性. 鉴于此，我们决定把它翻译成中文出版.

参与本书翻译工作的人员有：杨昔阳(福建泉州师范学院)、孙秋艳(北京市通州区潞河中学分校)、武方敏(厦门外国语学校海沧附属学校)、郭永丽(南京第二十七高级中学)、唐娟(北京三十五中)、邢岩(北京回民学校)、蔡瑞琼、张明欣等，它们对本书的部分章节作了翻译. 另外，蔡瑞琼、张明欣、李铭璜、俞盛利、曹玲玲、李洋等参与了本书的部分校对工作. 对他们所作出的辛勤劳动表示由衷的谢意. 特别要感谢北京师范大学黄崇福教授，他对本书的译稿进行了认真审阅，提出了很好的建议.

从策划选题，到与原书作者和有关出版社联系获得翻译出版权，再到翻译、校对、印刷、出版等历时 2 年，在这个过程中，北京师范大学出版社给与了全力支持，特别是王松浦老师为此付出了辛勤的劳动，她出色的策划工作使得本书能够顺利及时出版，在此表示衷心感谢. 同时，也一并感谢所有为此书的翻译及出版提供帮助的人士.

本书的翻译工作得到了国家自然科学基金资助(项目批准号：60775032)和北京师范大学校级重点学科资助.

由于译者水平有限，翻译中难免有不当或错误，请不吝赐教，谢谢.

于福生

北京师范大学数学科学学院

数学与复杂系统教育部重点实验室

2008 年 7 月 28 日

原著序言

给 Pedrycz 教授的专著写序言一直是个具有挑战性的任务，原因是，相对于现有文献中能找到的，他所写的已经前进了很远. 此次为书《基于知识的聚类：从数据到信息粒》简称《基于知识的聚类》写序更是如此. 《基于知识的聚类》是部杰作，它触及了一些人类认知的最基本层面，具有权威性、原创性、广博性、深刻性、和很强的解释性. Pedrycz 教授的书具有丰富的例子、图表和参考文献，这让人感觉读他的书是一件快事.

在这本书里，Pedrycz 教授谈及了一系列相关论题. 从解释聚类和模糊聚类开始，转向粒信息计算，其中的粒子是一簇属性值，这些属性值由于不可分辨性、等价性、相似性、相近性、或者功能性而结合在一起. Pedrycz 教授近来合作的一本有关粒计算的书为他提供了一个有效的框架，来把聚类和粒计算联系起来. 在粒计算中，运算的对象是粒子而不是点. 在它的一般形式中，粒计算包括区间计算、粗糙集合计算、概率分布计算. 粒计算和聚类分析的联结在 Pedrycz 教授的工作中扮演着关键角色，并且是他所提出的聚类分析方法的一个重要新颖的特征.

本书的核心是基于知识的聚类，其基础是聚焦于粒计算的那些章节. 在这种模式的聚类中，聚类被位于数据中的知识所指导. 在本书的这一部分里，特别是在研究条件模糊聚类、协作聚类、方向聚类、模糊关联聚类以及各向异性模式的聚类的章节里，有许多新颖的内容. Pedrycz 教授的工作的最后一个部分，可以说是基于知识的聚类应用于常用模型的一个增加见识的展览. 在这一部分里，我们能够发现一系列非传统的概念和技术，在它们当中有超盒建模、语言建模和粒映射.

为了从合适的角度看到 Pedrycz 教授工作的重要性，我作如下考察：随着我们进一步进入机器智能和自动推理时代，一个令人畏缩的问题变得越来越难以把握. 我们怎样才能应付数据、信息和知识的爆炸性增长？我们怎样才能从根植于一个无结构、不精确和非完全可信的大型数据库的决策相关信息中进行定位和推理？

这些问题所指向的，是在数据、信息和知识组织领域对新思想和新技术

的迫切需要. 事实上，信息是有组织的数据，而知识是有组织的信息.

组织概念中的一个关键概念是与关联性相关的概念，确切地说，是聚类和粒化的概念. 在这方面，作为 Pedrycz 教授工作研究对象的概念、思想和方法与设计一种能够处理数据、信息和知识的爆炸性增长的组织结构直接相关.

我还愿意作另一个考察. 尽管有大量文献涉及聚类分析和相关主题，但相当奇怪的是，在这些文献中我们找不到类概念的操作性定义. Pedrycz 教授的工作对类的有效性概念只作了简单的讨论，但却没有类概念定义的探讨.

是一个忽略，还是在定义类时存在一些问题需要解决？在我看来，有两个基本的问题. 首先，类概念是一个与度有关的概念，在这个意义下，类概念是一个模糊概念. 其次，一个类是点的集合而非单个点，在这个意义下，类概念是个二阶概念. 二阶概念的例子有凸集、边缘和一座山. 事实上，类的概念和山的概念具有相同的深层结构.

本质的问题是，二阶模糊概念通常不能在二值逻辑的概念框架中进行定义. 这也是在聚类分析的文献中找不到类的操作性定义的主要原因. 这里出现了一个问题：如果类概念不能在二值逻辑的概念框架中进行定义，那该如何定义它呢？在我看来，这需要 PNL(Precisiated Natural Language)—— 这种语言基于模糊逻辑——在模糊逻辑里，任何事物都是或都能被认为是与度有关的. 将 PNL 作为定义语言，用 PNL 定义一个概念，需要两个步骤. 首先，概念在自然语言中定义；其次，这种自然语言定义是精确的. 与基于二值逻辑的概念不同，基于 PNL 的定义是上下文相关而非上下文无关. 这是获得和实际更为一致的效果所必须付出的代价.

尽管在 Pedrycz 教授的工作中以至于在其他文献中没有类的定义，但这丝毫不降低他的工作的重要性. 对于每个感兴趣于聚类分析的人，或更一般地，对于每个感兴趣于基于知识的优秀系统的概念、设计和使用的人，《基于知识的聚类》都是必读的大作.《基于知识的聚类》是最杰出的工作，它的作者 Pedrycz教授和它的出版社 John Wiley& Sons，Inc. 理应得到我们的称赞和祝贺.

Lotfi A. Zadeh

原著前言

数据和模式是我们信息社会文化结构的必要组成部分. 我们每天面对的挑战是应付大量的数据，这些数据来自银行交易、成千上万传感器、万维网日记记录、单元呼叫通信通讯、卫星图形采集系统和智能家庭设备网络，这只说出了一些常见的例子.

弄懂数据的含义已变成智能数据分析、数据挖掘、传感器融合、图像理解和基于逻辑的系统建模的主要目标. 跟以前完全不同，我们正面临着这样一个日益增长的需要，就是要建立一只强有力的计算机“眼睛”——它是一种以人为中心、善与人交互、对人敏感的计算机环境，这种环境能有助于我们理解数据并作出明智决策.

聚类是这种计算机眼睛所具有的良好表现行为之一，因能探入数据空间并发现其中的数据结构——数据类，聚类成为一种理想的探知巨大数据空间的工具. 从 20 世纪 30 年代提出的早期概念开始，最近因受概念和计算的新挑战的刺激，这个领域发生了快速扩张. 现今，聚类的广泛存在是令人惊讶的. 即使是一次快速且相当不复杂的网络搜索或者是简单的图书馆数据库检索，就会得到大量相关项，这些相关项揭示了从生物医学到市场、工程、经济、生物学、化学、军事、食品工程、财经和教育等相当宽广的应用领域.

聚类已经成为多套方法和算法的同义词，这些方法和算法几乎全是数据驱动的，而且这些方法和算法中的任何一种优化，如果不是全部，也是绝大多数面向数据的. 聚类能产生一些用于揭示数据结构的信息粒. 粒化计算有助于设计聚类方法，以满足用户定义的目标. 在聚类的不同的风景图中，工作于模糊集框架内的算法取得了重要且独特的位置，原因很简单：被看作基本信息粒的模糊集是以人为中心的. 能够处理许可部分隶属的概念和组(类)使得模糊聚类具有高度吸引力. 能够识别具有边界特征、需特别关注进而确定为孤立点的数据是模糊聚类的一个有用、增值的特性. 能够发现类中最典型的模式(具有最大隶属度)是模糊聚类的又一重要特性.

根据代理技术、网络研究的最新应用和形式，以及快速扩张的数据维数及种类，聚类中以人为中心这一要求变得更加不可缺少，以数据为中心的聚

类必须进行扩张．我在本书介绍的基于知识的聚类涉及如何协调聚类活动中的两个重要驱动力：获取数据和领域知识，在高维且通常不同质的数据空间中建立一个一致的研究平台．任何一个高度交互的数据分析在形成一个不可缺少的反馈式循环的过程中，用户扮演着重要的角色．更不用说，我们需要一个精心挑选的用于人机交流的概念和算法层．

这本书分为三部分，第一部分包括第一章到第三章，提供了精炼、认真构建的主题介绍，呈献了三个相关联的部分：第一，讨论了模糊聚类的基础知识．第二，从模糊聚类的层面对被视为粒化计算重要实现形式之一的模糊计算作了回顾．第三，对基于逻辑的神经元及相关联的神经网络作了详细介绍．第四章到第十章是这本书的中心内容，提供了基于知识聚类的极其不同的风景介绍．这本书的第三部分包括第十一章到第十五章，专注于一些一般的模型，这些模型的设计与基于知识的聚类直接相关．首先，专注于类的超盒模型，展示如何根据超盒的几何特性来捕捉本质结构，紧接着，研究了粒映射和语言模型．

纵贯全书，我坚持使用模式识别和系统分析中的标准符号，以及那里使用的术语．词语“数据”和“模式”的交替使用是为了强调处理各种形式的模式识别、系统建模和数据分析方法的统一性．这本书是自身完备的，尽管读者能够从初步熟悉计算智能中受益，但这不是本书必需的东西．计算智能有助于透视资料，且能让读者全面鉴赏作为计算智能结构构件的信息粒度和信息粒．

本书的目的是将一些主要的思想以极一般的形式呈献出来，而且不会因为将讨论局限于一些选取的应用领域而偏离主题．在算法方面也是保持了一般性探讨，并没有着力去苛求那些可能最有效但繁琐的实现方法，这使得更广大的读者对本书有兴趣．那些对聚类、模糊聚类、无监督学习、神经网络、模糊集和模式识别感兴趣的读者，以及潜心于各种数据分析任务的人，会发现本书是引人思考和启发智力的．致力于系统建模的读者将把知识驱动的聚类看作是快速建立粒模型的富有吸引力的工具．

基于知识的聚类已经出现，这本书概括性地论述了它的基础知识，呈献了重要的算法设计，并对其应用驱动的层面进行了讨论．虽然没有着力去完全涵盖主题的所有方面，但是，所选取的材料为在这一快速发展的领域中居于中心的最新进展描绘了一幅条理清晰的图画．

Witold Pedrycz

目　录

第 1 章　聚类和模糊聚类

这一章对聚类作全面聚焦的介绍，聚类是探测性数据分析、无监督学习、数据粒化以及信息压缩的基本工具. 我们讨论多种聚类算法(包括分层算法和基于目标函数的算法这样一些基本算法)的基本原理，详细说明它们的分类，并回顾与这些聚类算法相关联的解释机制.

1.1　引言

让数据变得有意义是研究人员和专业人员在几乎所有现实努力中所面临的任务. 以海量数据为特征的信息科技的时代更是放大了这一需求并使之变得更具挑战性. 时时、处处都在产生数据聚集已经成为我们生活中的现实. 理解这些数据、揭示其中的基本现象、可视化其主要的趋势是智能数据分析(IDA)、数据挖掘(DM)和系统建模要努力完成的主要任务.

聚类是用于数据分析和解释的一般方法和明显富于概念特征和算法特征的框架(Anderberg，1973；Bezdek，1981；Bezdek 等，1999；Devijver 和 Kittler，1987；Dubes，1987；Duda 等，2001；Fukunaga，1990；Hoppner 等，1999；Jain 等，1999，2000；Kaufmann 和 Rousseeuw，1990；Babu 和 Murthy，1994；Dave，1990；Dave 和 Bhaswan，1992；Kersten，1999；Klawonn 和 Keller，1998；Mali 和 Mitra，2002；Webb，2002). 在这一章里，我们介绍基本概念，解释聚类问题公式中重要的功能成分，并讨论聚类算法的主要类别. 这些算法伴随着粒计算的形式，包括集合、模糊集、阴影集和粗糙集.

1.2　基本概念和符号

为了建立一个形式框架，使得在其中能实施聚类，我们先讨论一些基本概念，包括数据类型、距离，以及相似性/相像性等.

1.2.1　数据类型

我们身边的世界产生了大量各种类型的数据，数据形式的多样性给人印象深刻. 模式的形式表示和组织反映了我们打算处理数据的方式. 常用的最一般的分类分为数值的(连续的)、有序的和名词性的变量. 一个数值变量可以取实数域 **R** 的任意值；一个有序变量可以取少量的离散状态值，并且这些

状态是可以比较的. 例如，有四个状态，记为 a_1，a_2，a_3 和 a_4，我们可以说 a_1 离 a_2 比 a_1 离 a_3 来的近；一个名词性变量可以取少量状态值，但是却无法谈论这些状态的远近. 除了这点区别，有序变量和名词性变量都被表示为离散的变量. 为了进行计算，我们通常有几种编码机制可以采用，比方说二进制编码和带有不同选项的二进制编码.

变量可以被组织到能够反映问题特点的内在结构中去. 如果每个模式都是被一些特征所描述，那么我们会非常直观地把它们写成一些向量，比方说 $\boldsymbol{x}$，$\boldsymbol{y}$，$\boldsymbol{z}$. 根据所使用变量的特点，向量的分量可以是实数或者二进制数. 显然，这会导致各种形式的向量，这些向量包含了两种类型的分量. 作为特征向量的分量，向量和矩阵中的所有变量具有相同的层次，并且没有结构，从这个意义上讲，向量和矩阵是“平凡”的结构. 像树一样的层次结构则用于将聚类或分类过程中我们感兴趣的对象(模式)之间的关系形象化.

1.2.2 距离和相似性

不相似性(或距离)的概念或者对偶的相似性概念是任何一种形式的聚类的重要组成部分，它能帮助我们穿过数据空间并形成类. 通过计算不相似性，我们能理解并清楚地说出两个模式有多靠近，并根据这个靠近程度，把它们分派到同一个类. 正式地说，$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的不相似性 $d(\boldsymbol{x}, \boldsymbol{y})$可以看作是一个满足以下条件的二元函数：

$$\begin{aligned} &d(\boldsymbol{x},\boldsymbol{y}) \geqslant 0 \quad \text{对每个 } \boldsymbol{x} \text{ 和 } \boldsymbol{y}, \\ &d(\boldsymbol{x},\boldsymbol{x}) = 0 \quad \text{对每个 } \boldsymbol{x}, \\ &d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{y},\boldsymbol{x}), \end{aligned} \tag{1.1}$$

这些要求直觉上很吸引人. 我们需要不相似性的非负性质，对称性也是个显然的要求. 不相似性在处理两个相同模式时达到全局最小值，即 $d(\boldsymbol{x}, \boldsymbol{x})=0$.

距离(度量)，是个更具限制性的概念，因为我们要求满足三角不等式，即对任意的模式 $\boldsymbol{x}$，$\boldsymbol{y}$ 和 $\boldsymbol{z}$，我们有：

$$d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z}) \geqslant d(\boldsymbol{x},\boldsymbol{z}), \tag{1.2}$$

在连续特征(变量)的情形下，我们有许多距离函数(表 1.1). 这些函数中的每个都因为几何特性的不同而意味着对数据的不同理解. 当我们仅考虑两个特征($\boldsymbol{x}=[x_1\ x_2]^{\mathrm{T}}$)并计算 $\boldsymbol{x}$ 到原点的距离时，其几何特性很容易图示说明. 常距离的等高线(图 1.1)展示了什么样的几何结构成为搜寻结构的焦点. 这里，我们意识到 Euclidean 距离适合于具有圆周形状的数据类. 距离函数具有某个类别；Minkowski 距离包含无穷多距离函数，包括一些熟知和常用的距

离，像 Hamming，Tchebyschev 和 Euclidean 距离.

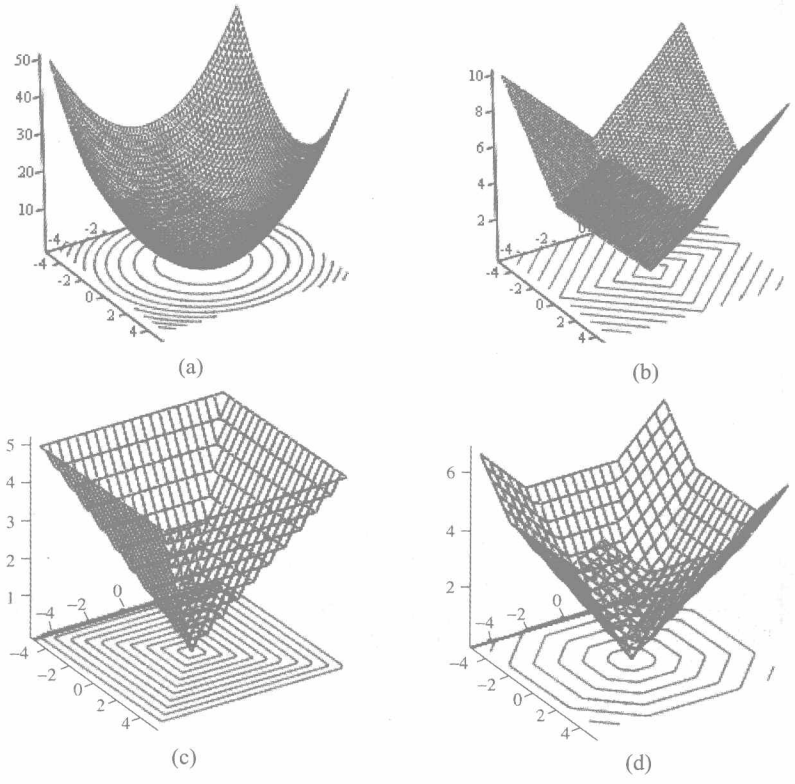


图1.1　距离函数示例——三维和等高线图：(a) Euclidean，(b) Hamming (city block)，(c) Tchebyschev，(d)组合距离 max (2/3 Hamming，Tchebyschev).

表 1.1　选取的关于模式 x 和 y 的距离函数

距离函数	公式及注释
Euclidean 距离	$d(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
Hamming(city block) 距离	$d(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}\lvert x_i-y_i\rvert$
Tchebyschev 距离	$d(\boldsymbol{x},\boldsymbol{y})=\max_{i=1,2,\cdots,n}\lvert x_i-y_i\rvert$

续表

距离函数	公式及注释
Minkowski 距离	$d(\boldsymbol{x},\boldsymbol{y})=\sqrt[p]{\sum_{i=1}^{n}(x_i-y_i)^p}$，$p>0$
Canberra 距离	$d(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}\frac{\lvert x_i-y_i\rvert}{x_i+y_i}$，$x_i$ 和 y_i 均为正数
角分离	$d(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}x_iy_i\Big/\Big[\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2\Big]^{1/2}$ 注：这是一个表达 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的单位向量之间夹角的相似性测度.

当 Minkowski 距离函数中的指数变化的时候，能达到图 1.1(d)展示的相同效果，见图 1.2.

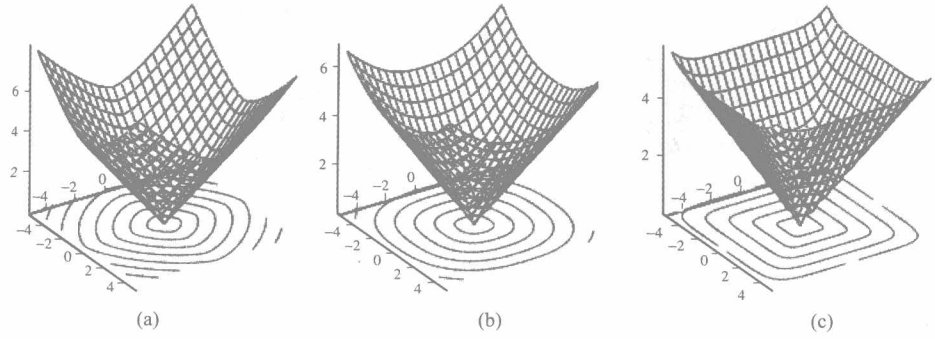


图1.2 带有不同指数的 Minkowski 距离函数的例子：(a) 1.5 (b) 2.5，和(c) 7.0.

一个常用的一般形式是 Mahalanobis 距离

$$d(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{y}, \tag{1.3}$$

其中，$\boldsymbol{A}$ 是正定矩阵. 通过选取这个矩阵，以及旋转椭圆(离开 A 的对角线元素)和改变坐标轴的长度(位于矩阵主对角线上的元素)，我们能控制潜在类的几何特性.

对于二进制变量，我们通常聚焦于相似性而不是距离(或不相似性). 考虑两个由二进制数据串$[x_k]$，$[y_k]$组成的二进制向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$；两两比较它们的对应分量，对发生的情形进行简单计数：

x_k 和 y_k 都等于 1 发生的次数，

$x_k=0$ 和 $y_k=1$ 发生的次数，