K-means 后数据聚类的 50 年发展

Anil K.Jain 密歇根州立大学计算机科学与工程系 高丽大学大脑与认知工程系

翻译人 徐天宇 专业班级 自动化 1104 .

摘要：数据进行合理的聚群是理解和学习最基本的模式之一。例如，一个常见的科学分类将生物归类为如下的类别体系：域、界、门、纲、目等。聚类分析是根据对象的可测得的或可感知的本质特征或相似度来对其进行聚群或聚类的方法和算法的正式研究。聚类分析并不使用种类标签，即通过如类标这样已有的标示符来标识对象。类别信息的缺失将数据聚类（无监督学习）和分类或判别分析（有监督学习）。聚类的目标是寻找数据的结构，因此是对自然的一种探索。聚类在不同的科学领域里面都有着悠久而丰富的历史。1955 年第一次发表的 K-means 算法是最受欢迎的简单聚类算法之一。事实上，尽管 K-means 算法已经提出了 50 多年，而且从那时起发表了数以千计的其它聚类算法，K-means 仍然有着广泛的运用。这说明设计一个有广泛适用性的聚类算法的困难以及聚类本身是一个病态问题。我们对聚类进行了简要的综述，总结了有名的聚类方法，讨论了设计聚类算法主要挑战和核心问题，指出了部分新兴和有用的研究方向包括半监督聚类、集成聚类、在数据聚类时同时进行特征选择以及大规模数据聚类。

关键词：数据聚类、用户困境、历史发展、聚类的前景、傅京孙奖

1. 引言

传感和存储技术的进步以及像互联网搜索、数字成像、视频监控等技术应用的迅猛发展产生了大量的高维数据集。据估计 2007 年数据全球数据使用量为 281 艾字节，预计 2011 年这个数字将增长 10 倍（1 艾大约是 10^{18}B 或 1,000,000TB）。

大部分的数据数字化的存储在电子介质中，因此给自动化数据分析、分类和检索技术的发展提供了巨大的可能。可利用的数据除了量的增长，类型也增多了（文本、图像、视频）。并不昂贵的数字摄影机产生了大量的图像和视频。由于无线射频识别标签和收发机低价和小尺寸，它们得以普及并导致了成千上万的能有规律传输数据的传感器的部署。E-mail、博客、交易数据以及数以亿计的网页每天产生数 TB 的新数据。很多这类数据流都是松散的，给分析它们增加了难度。

数据数量和类别两方面的增长迫切的需要自动理解、处理和概括数据的方法的进步。数据分析方法可以概括的分为主要的两类（Tukey,1997）:(i)探索性的或描述性的，指研究者没有事先明确的模型或假设但是想理解高维数据的大体特征和结构。（ii）验证性的或推理性的，指研究者想要验证适用于可用数据的假设/模型或一系列假定。很多统计学方法被用来分析数据，举几个例子，比如方差分析、线性回归、判别式分析、典型相关分析、多维定标、因子分析、主成分分析和聚类分析。一个相关有用的综述已经发表（Tabachnick 和 Fidell，2007）。

在模式识别中，数据分析涉及预测建模：给定一些训练数据，我们想要预测未知测试数据的行为。这个任务也被叫做“学习”，通常两类学习之间有明确的区别（i）有监督的（分类）。（ii）无监督的（聚类）。第一种只涉及有标签的数据（训练模式有已知的类别标签），而第二种只涉及无标签数据（Duda et al.2001）.聚类相比分类是一个更加困难更有挑战性的问题。一种混合的设置，即半监督学习正在受到越来越多的关注（Chapelle et al. 2006)。在半监督分类中，训练数据集中只有一小部分的标签是可用的。那些没有标签的数据，也在学习过程中使用，而不是被放弃了。在半监督聚类中，有些对间约束是明确的，而不是有明确的类标，也就是有一种弱化的先验知识。一个对间“必须连接”相当于要求两个对象被赋予相同的聚类标签，反之，共享一个“不能连接”约束的两个对象的聚类标签应该是不同的。在潜在簇的准确定义缺失的数据聚类中，约束可以变得非常有用（Lange et al.,2005;Basu et al.,2008）。为了寻找好的模型，我们会应用所有的可用信息，不管是否是无标签的数据、有约束的数据还是有标签的数据。图 1 举例说明了这些模式识别和机器学习所感兴趣的不同类型的学习问题。

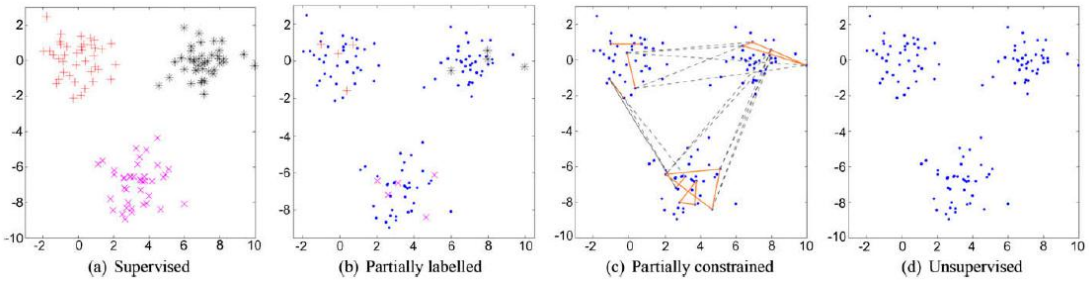

图 1. 学习问题：圆点表示不带有标签的点。带有标签的点由加号、星号和叉号表示。在图（c）中，必须连接和不能连接约束由实线和虚线各自表示（图片取自 Lange et al.(2005)）。

2. 数据聚类

数据聚类或者说聚类分析的目标是发现一系列模式、点或者对象的天然的分组情况。Webster（Merriam-Webster 在线字典,2008）将聚类分析定义为“关于通过定量比较多重特性发现群体中的个体是否属于不同的组别的一种统计分类方法。”图 2 是聚类的一个例子。目标是开发出一种可以从图 2a 中的无标签数据中发现图 2b 中的自然分组的自动化算法。

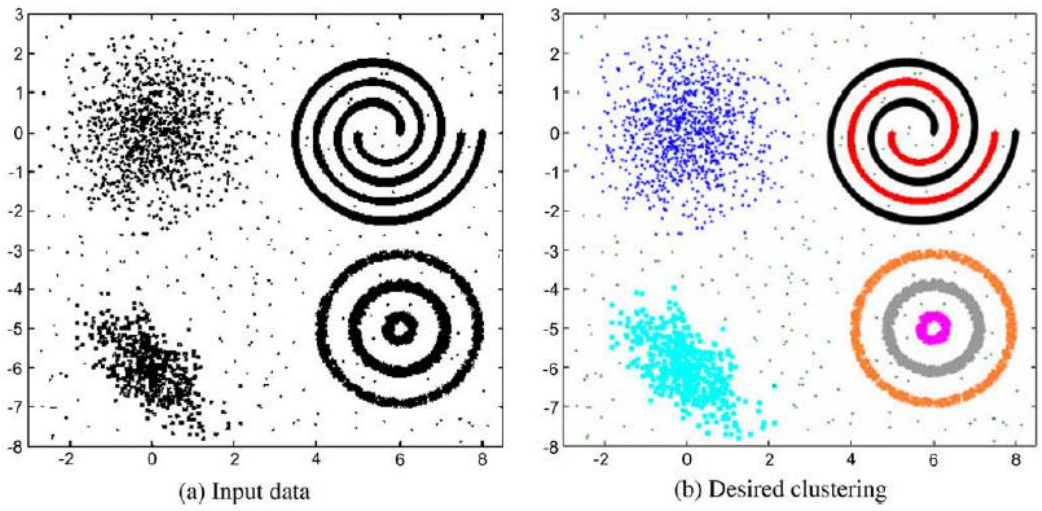

图 2. 各种各样的簇。图（a）（在图 1（b）中由 7 种不同的颜色表示）中的 7 个簇在形状、大小和密度上都不同。虽然这些簇对数据分析师来说是显而易见的，但是目前还没有可用的聚类算法可以找出所有这些簇。

聚类的一种实用性定义可以表示如下：给定 n 个对象的某种表示，找到基于相似度测量的 K 个分组使得在同一个分组中的对象间相似度高而处于不同分组

中的对象间相似度低。但是，相似度的定义是什么？簇的定义又是什么？图 2 表明簇的形状、大小和密度可以是不同的。数据集中存在的噪声使得发现簇更加的困难。一个理想的簇可以被定义为一系列紧凑的、孤立的点集。事实上，从一个观看者的角度来看，一个簇是一个主观的实体，它的重要性和解释需要相应领域的知识。虽然人类本身是二维可能三维数据中簇的出色探求者，但是我们需要自动化算法来处理高维数据。聚类时因为不知道所给定数据到底可以分成几个簇，导致了数以千计的聚类算法的出现和将继续出现。

2.1. 为什么要聚类？

在任何涉及多变量数据分析的学科中聚类分析都是很普遍的。通过谷歌学术搜索引擎（2009）搜索关键词 “数据聚类”仅仅出现在 2007 年的就有 1660 个条目。如此大量的文献足以说明聚类在数据分析中的地位。很难一一列举聚类方法在众多科学领域和应用领域的使用，同样，也很难一一列举已经发表的数以千计的聚类算法。图像分割作为计算机视觉中的一个重要问题可以表示为一个聚类问题（Jain 和 Flynn，1996；Frigui 和 Krishnapuram，1999；Shi 和 Malik，2000）。文档可以被聚类（Iwayama 和 Tokunaga，1995）为几个有效信息访问或检索的局部层次结构（Bhatia 和 Deogun，1998）。聚类被用来将顾客聚群为不同的类型以便进行有效的营销（Arabie 和 Hubert，1994）；被用来聚群服务提供协议以便进行劳动力管理和规划（Hu et al.,2007）；还被用来研究生物学中的基因组信息（Baldi 和 Hatfield，2002）。

数据聚类主要被用在以下三个地方：

(a)构造底层结构：获得对数据的深入了解，产生假设，侦测异常，识别特征。

(b)自然分类：识别形式或生物间的相似程度（系统发育关系）。

(c)压缩：通过聚类原型作为一种组织和概括数据的方法。

图 3 是类别发现的一个例子。这里聚类被用在一个在线手写字符识别应用程序中来发现不同的子类（Connell 和 Jain，2002）。不同的用户会用不同的方式写同一个数字，所以要适当增加类内差异。在一个类中聚类训练模式可以发现新的子类，叫做手写字符的词位。使用基于不同子类的多重模型而不是每个字符的单一模型可以用来提高识别的准确率（见图 3）。

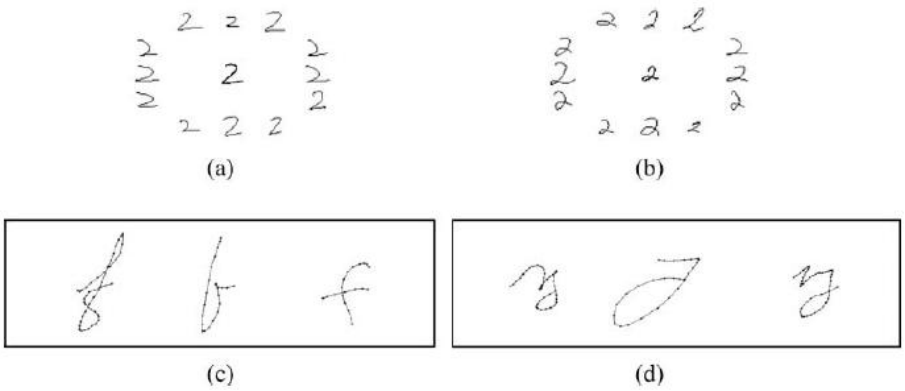

图 3. 使用数据聚类找到子簇。图（a）和（b）以两种不同的方式写数字 2；图（c）是字符‘f’的三种不同的子类；图（d）是字母‘y’的三种不同子类。

给定大量的英特网上的网页，很多查询词条会典型的返回大量的网页可点击数。这产生了对组织搜索结果的需要。像 Clusty（www.clusty.org）这样的搜索引擎聚类了搜索结果，并将一个更加有序的结果呈现给用户。

2.2. 历史发展

聚类方法的发展确实是多学科共同努力的结果。分类学家、社会科学家、心理学家、生物学家、统计学家、数学家、工程师、计算机科学家、医学研究者以及其他收集和处理真实数据的人都为聚类方法做出了贡献。根据 JSTOR（2009），数据聚类第一次出现是在 1954 年发表的一篇处理人类学数据的文章的标题中。数据聚类根据不同的的应用领域也被叫做 Q 分析、类型学、聚丛、分类学（Jain 和 Dubes，1988）。目前有几本关于数据聚类的书，经典的是由 Sokal 和 Sneath（1963）、Anderberg（1973）、Hartigan（1975）、Jain 和 Dubes（1988）和 Duda et al.(2001)写的。数据挖掘领域也广泛研究了聚类算法（参见 Han 和 Kamber（2000）和 Tan et al.(2005)写的书以及《机器学习》（Bishop，2006））。

聚类算法可以粗略的分为两组：基于层次的和基于划分的。基于层次的聚类算法使用凝聚的模式（从将每个数据点作为一个簇开始相继的融合最相似的一对簇为一个新的簇层）或者分裂（自顶向下）的模式（从将所有数据点作为一个簇开始递归的分裂每个簇为一个更小的簇）递归的找到嵌套的簇。和基于层次的聚类算法相比，基于划分的算法找到所有的簇的同时也找到了数据的一个划分而且并不使用层级结构。基于层次的算法的输入是一个 n*n 的相似度矩阵，期中 n

是用于聚类的数据对象的数量。另一方面，基于划分的算法既可以使用 n*d 的模式矩阵，期中 n 是表示有 d 维特征空间的 n 个对象，也可以使用相似度矩阵。值得注意的是相似度矩阵可以很容易的由模式矩阵导出，但是要从相似度矩阵中导出模式矩阵就要使用像多维定标（MDS）这样的定标方法。

最著名的基于层次的算法是单一连接和完全链接；最受欢迎也最简单的基于划分的算法是 K-means。由于可用数据本身的性质决定基于划分的算法在模式识别中更加受欢迎，我们在这里主要讨论这类算法。K-means 自从在不同的科学领域中由 Steinhaus（1956）、Lloyd（1957 年提出，1982 年发表）、Ball 和 Hall（1965）以及 MacQueen（1967）独立的发现以来，已经有了丰富而多样的历史。即使 K-means 距离它第一次提出已经过去了 50 多年，它仍然是聚类中最广泛使用的算法之一。易实现、简单、有效、经验上的成功是这个算法如此受欢迎的主要原因。下面，我们先总结一下 K-means 的发展历程，然后讨论数据聚类中的主要的成熟方法。

2.3. K-means 算法

令 $X=\{x_i\}$，i=1，...,n 是要用来分成 K 个簇 $C=\{c_k$，k=2，...,K}的 d 维点集。K-means 算法找到一个使得一个聚类中经验均值和点之间的平方误差最小。令 μ_k 为 c_k 的均值，类 c_k 中 μ_k 和点之间的平方误差定义为 $J(c_k)=\sum_{x_i\in c_k}\|x_i-u_k\|^2$。K-means 的目标是使得所有 K 个聚类中平方误差的和 $J(C)=\sum_{k=1}^{K}\sum_{x_i\in c_k}\|x_i-u_k\|^2$ 最小。最小化这个目标函数是一个 NP 困难问题（即使 K=2）（Drineas et al.,1999）。因此 K-means 是一个贪心算法，只能收敛于局部最优解，即使最近的研究表明当簇很好分开时 K-means 有很大的可能概率收敛于全局最优解（Meila，2006）。K-means 从初始划分的 K 个聚类开始，并给簇分配模式以便减少平方误差。因为当簇数目 K 增加时平方误差总会减少（当 K=n 时，J（C）=0），只有当聚类数目是固定数量时才可以最小化 J（C）。K-means 算法的主要步骤如下（Jain and Dubes，1988）：

1.选择一个有 K 个簇的初始划分；重复步骤 2 和 3 直到每个聚类中的对象稳定。

2.通过将每个模式分配给各距离簇中心最近的簇产生一个新的划分。

3.计算新簇中心。

图 4 是一个有三类的 2 维数据集 K-means 算法的例子。

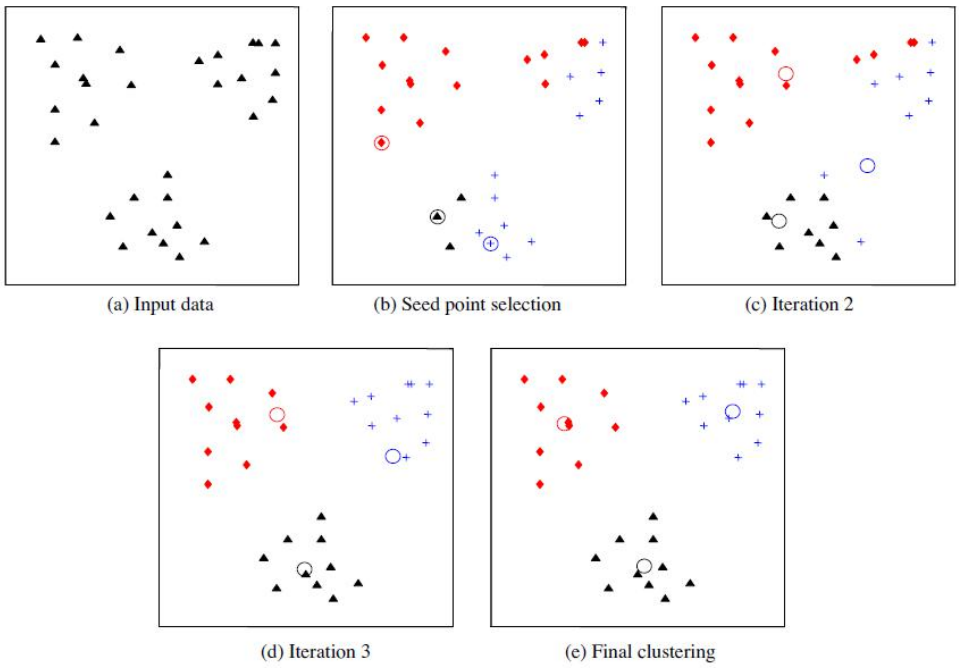

图 4. K-means 算法的例子。图（a）表示有三个簇的二维输入数据；图（b）表示了用作簇中心的三个种子点以及数据点的初始簇分配。图（c）和（d）表示簇标和簇中心的中间迭代。图（e）表示 K-means 收敛最后得到的聚类结果。

2.4. K-means 的参数

K-means 算法需要用户确定三个参数：类的数目 K、类的初始化、距离尺度。其中最重要的参数是 K。因为目前还没有准确的数学标准存在，可以选择已有的许多试探性求解（见 Tibshirani et al.，2001）K 的方法。一般来说，K-means 在不同的 K 下会独立的运行，并产生不同的划分，对于领域专家来说最有意义的那种划分就会被选择。因为 K-means 只能收敛于局部最优解，不同的初始化会产生不同的聚类结果。一种寻找全局最优解的方法是在一个给定的 K 下，使用不同的初始划分多次运行 K-means 算法，然后选择平方误差最小的划分。

K-means 一般会选择欧几里得距离作为距离尺度来计算点和簇中心之间的距离。因此 K-means 找到的是数据中球面和球体的簇。K-means 利用马氏距离可以被用来寻找超椭球体的簇（Mao 和 Jain，1996），但是这需要更大量的计算花费。一个 K-means 的变体使用板仓距离来处理在语音处理中的矢量量化问题

（Linde et al.，1980）。L_1 距离被建议用来为 K-means 开发 Bregman 距离族（Kashima et al.,2008.Banerjee et al.2004）。

2.5. K-means 的扩展

基础的 K-means 算法已经被很多不同的方式扩充了。有些扩充使用额外的试探性方法处理簇尺寸最小化，合并和分离簇。在模式识别文献中的两个有名的 K-means 变体是 ISODATA(Ball 和 Hall，1965)和 FORGY（Forgy，1965）。在 K-means 中，每个数据点被分配到一个单一簇中（叫做硬分配）。由 Dunn（1973）提出，并由 Bezdek（1981）改进的模糊 c-means，作为 K-means 的一个拓展，它的每个数据点可以是多个簇的成员（软分配），其中用成员值加以区分。一个好的可供使用的关于基于模糊的聚类综述是（Backer，1978）。在聚类之前，通过用分组数据的中心代替原来的数据可以加速 K-means 和模糊 c-means（Eschrich et al.,2003）。下面概括 K-means 的其它重要的修正算法：Steinbach et al.(2000)提出了一种 K-means 的层次分裂版本，叫做二等分 K-means，它每一步迭代划分数据为两个簇。在（Pelleg 和 Moore，1999）中，kd 树被用来处理 K-means 的核心步骤，即有效的识别所有数据点中最近的簇中心。Bradlry et al.(1998)提出了 K-means 的一个快速可伸缩单次扫描版本，它不需要所有的数据同时适合内存。X-means（Pelleg 和 Moore，2000）可以通过优化比如 AIC 或 BIC 这样的准则自动化的找到 K。在 K-medoids（Kaufman 和 Rousseeuw，2005）中，由数据的中位数代替均值表示簇。核心 K-means（Scholkopf et al.,1998）被提出来通过选择核心相似度函数用于发现任意形状的簇。值得注意的是所有这些拓展都要增加由用户确定的算法附加参数。

2.6. 聚类的主要方法

就像之前提到的一样，在很多不同的学科的文献中提出了数以千计的聚类算法。这使得回顾所有这些已经发表的方法变得非常困难。好在，聚类算法的主要差异是在目标函数的选择、概率生成模型和试探法三个方面。我们将简要回顾一些主要的方法。

簇可以被定义为在特征空间中由低密度区域分开的高密度区域。算法根据这

个簇的定义直接搜索特征空间中连通的稠密区域。不同的算法使用不同的连通性定义。Jarvis-Patrick 算法定义一对点之间的相似度为它们共享的邻近点的数量，临近点是指出现在某一点的一定半径区域内的点（Frank 和 Todeschini，1994）。Ester et al.(1996)提出了和 Jarvis-Patrick 算法相似的 DBSCAN 聚类算法。它通过使用 Parzen 窗方法估计密度，直接搜索连通的稠密区域。Jarvis-Patrick 算法和 DBSCAN 的性能依赖于两个参数：距离角度的邻域尺寸和一个簇的邻域内包含的点的最小数量。另外，许多概率模型被开发用于数据聚类，即通过混合概率模型模拟密度函数。这些方法假设数据由一个混合分布生成，即每个簇被一个或多个混合分布的组合来描述（McLachlan 和 Basford，1987）。EM 算法（Dempster et al.,1977）经常被用来推断混合模型的参数。包括潜在狄利克雷分布（LDA）（Bleiet al.2003）、弹球分布模型（Li 和 McCallum，2006）和无指导数据聚类图形模型（Welling et al.2005）在内的几个贝叶斯方法已经被开发来改进用于数据聚类的混合模型。

虽然基于密度的方法，尤其是无参数的基于密度的方法因为可以处理任意形状的簇的内在性质而很有吸引力，但是它们在处理高维数据时却有局限性。当数据是高维的，特征空间往往是高维的，这使得从低密度区域中区分高密度区域变得很难。子空间聚类算法通过找到嵌入在低维子空间的给定高维数据的簇来克服这个局限性。CLIQUE（Agrawal et al.,1998）是一个可伸缩的聚类算法，设计用于寻找数据中有高密度簇的子空间。因为它只在低维度子空间进行估计，CLIQUE 并不会遇到高维问题。

图论聚类，有时也被叫做谱聚类，用一个加权的图代表数据点。连接两个节点之间的边由对间相似度加权。主要思想是：划分节点为 A 和 B 两个子集使得剪裁尺寸，也就是被分配给连接节点集 A 和 B 的边的权值的和是最小的。最初用于解决这个问题的算法是最小剪切算法，它经常导致两个簇有不平衡的尺寸。后来比率剪切算法采用了一种簇尺寸（一个簇中的数据点的数量）约束（Hagen 和 Kahng，1992）。另一种叫做规范化剪切的带有簇尺寸（簇的数据量或者单个簇内的边的权值和）的近似基于图剪切的有效聚类算法是由 Shi 和 Malik（2000）第一次提出的。它的多级版本由 Yu 和 Shi（2003）提出。Meila 和 Shi（2001）提出了谱聚类的一种马尔科夫随机漫步观点并提出了可以处理任意簇数目的修

正规范化剪切（MNCut）算法。Ng et al.(2001)提出了另一种谱聚类算法的变体，它从核心矩阵的规范化特征向量中导出一个新数据表示方法。拉普拉斯特征映射（Belkin 和 Niyogi，2002）是另一种谱聚类方法，它基于图的拉普拉斯算子导出数据表示方法。Hofmann 和 Buhmann（1997）为聚类提出了一种确定性退火算法，它利用数据对象间的邻近衡量来表示数据。Pavan 和 Pelillo（2007）通过将最大支配集（Motzkin 和 Straus，1965）和簇相联系明确的表达了对间聚类问题，它是一个图中簇的一种连续产生方法。

有些聚类算法有一个信息理论公式。比如，Roberts et al.(2001)提出的最小熵方法假设数据是由一个混合模型生成的并且每个簇是使用一个半参数概率密度模式化的。参数通过最大化簇中数据点的无条件密度和有条件密度之间的 KL 散度来进行估计。这使得有条件和无条件的密度之间的重叠最小化，因此将簇分隔开。换句话说，这个公式是利用最小化所观察数据的预期熵的方法的结果。信息瓶颈方法（Tishby et al.，1999）被提出来作为速率失真理论和应用有损耗数据压缩观点的一般化。简单的说，给定两个随机变量的联合分布，信息瓶颈在最大化保留两个变量间的相互信息的前提下，压缩其中一个变量。（Slonim 和 Tishby,2000）展示了这种方法在文档聚类中的应用，其中的两个随机变量是单词和文档。单词通过使得和文档之间的信息得以最大化的保留来聚类，使用聚类后的单词，文档通过使得聚类后单词和聚类后文档之间的相互关系得到最大化的保留来聚类。

3. 用户的困境

尽管有如此大量的聚类算法，而且这些算法也在许多不同的领域成功应用，聚类仍然是一个困难的问题。这可以归因于簇定义本身的模糊性以及定义一个合适的相似度尺度和目标函数。

（Jain 和 Dubes，1998）强调以下聚类的主要挑战，这即使到今天来看依然是中肯的。

（a）什么是一个簇？

（b）应该使用哪些特征？

（c）数据需要标准化吗？

（d）数据是否有异常值？

（e）怎样定义对间相似度？

（f）数据中有多少簇？

（g）应该使用哪个聚类方法？

（h）数据是否有聚类趋势？

（i）发现的簇和划分是否有效？

以下我们会强调和举例说明其中一些挑战。

3.1. 数据表示法

数据表示是影响聚类算法性能最重要的因素之一。如果表示法（特征的选择）很好，那么簇很有可能是紧凑而孤立的，即使是一个像 K-means 这样的简单聚类算法也能找到它们。可惜的是，没有通用的表示法，表示法的选择必须在领域知识的指导下进行。图 5a 表示的是一个 K-means 不能将其划分为两个自然簇的数据集。通过 K-means 得到的划分由图 5a 中的虚线表示。然而当 a 中同样的数据点利用数据计算得来的 RBF 相似度矩阵的顶层二维特征向量使其表示成图 b。它们分离的非常好使得用 K-means 来聚类这些数据非常简单（Ng et al.,2001）。

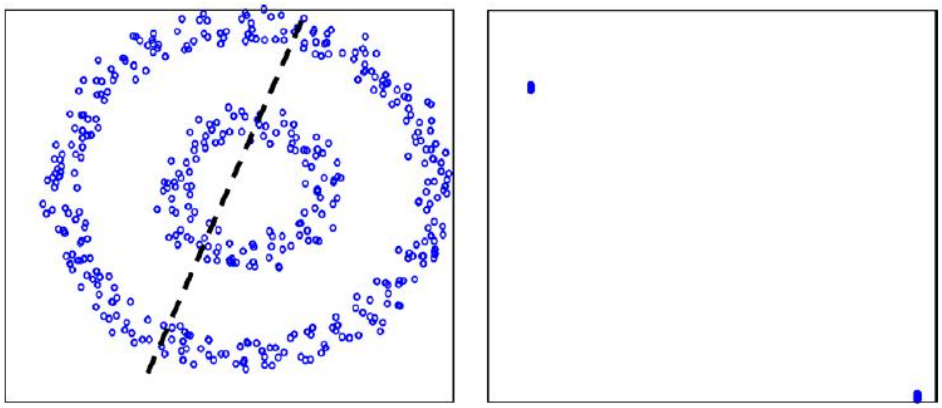

图 5. 一个好的表示法的重要性。图（a）表示 K-means 无法找出“自然”簇的“两环”数据集；虚线表示 K-means 在参数 $K = 2$ 情况下获得的线性簇分离边界；图（b）是使用 RBF 核心计算基于数据的图拉普拉斯算子的顶层两个特征向量的图（a）的一种新的表示法；K-means 现在可以简单的找到两个簇。

3.2. 分组的目的

数据的表示和分组的目的密切联系。表示法必须和用户最终的目的相配合。（Pampalk et al.,2003）中用 13 个布尔特征表示 16 种动物的例子证明表示法是怎样影响分组的。动物通过关于它们的外貌和活动的 13 个布尔特征表示。当更多的特征是外貌特征而不是活动特征时，动物被分成哺乳动物和鸟类。另一方面，当更多的特征是是活动特征时，数据集被分成食肉动物和非食肉动物。图 6 中的两个划分同样有效，它们都揭示了数据有意义的结构。要获得想要的聚类结果就要靠用户仔细的选择表示法。

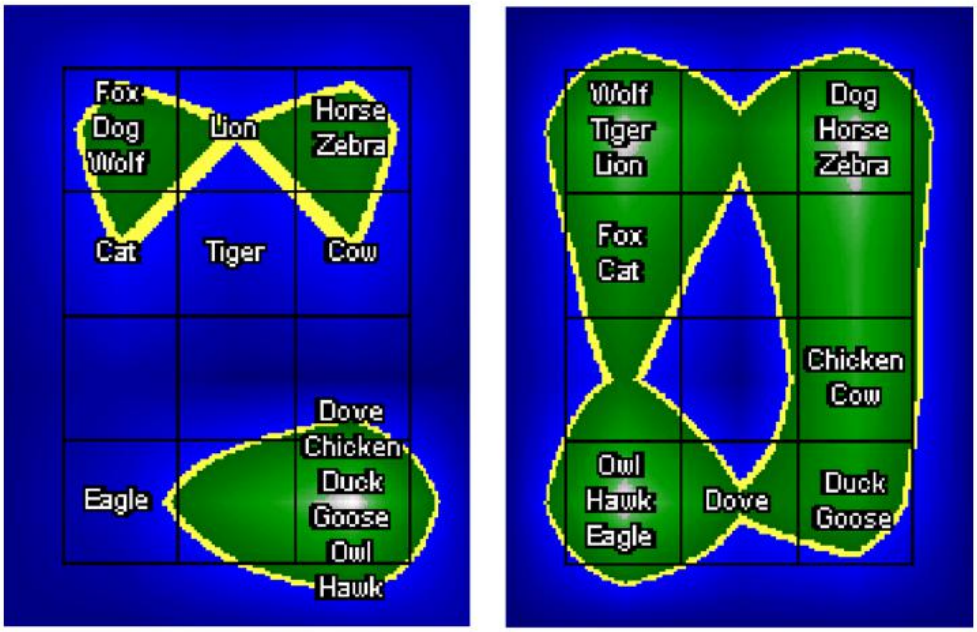

图 6. 数据特征的不同权重导致数据的不同划分。16 种动物由与外形和活动相关的 13 个布尔特征值表示。图（a）表示当基于外形的特征获得较大权重时的划分结果；图（b）表示当基于活动的特征获得较大权重时的划分结果。图（a）和（b）引用自 Pampalk et al.(2003), 被称之为“热图”，其中颜色代表一个位置的样本密度；颜色越暖，密度越大。

3.3. 簇的数目

自动确定簇的数目是数据聚类中最困难的问题之一。大部分自动确定簇数目的方法是将其转化为模型选择问题。通常，聚类算法在不同的 K 值下运行，然后根据之前定义的准则选择最佳的 K 值。Figueiredo 和 Jain（2002）使用最小信息长度（MML）准则（Wallace 和 Boulton，1968；Wallace 和 Freeman，1987）结合高斯混合模型（GMM）来估计 K。他们的方法从大量的簇开始，然后朝着

减少 MML 准则方向逐步融合已有的簇。一个使用最小描述长度（MDL）原理的相关方法被用在（Hansen 和 Yu，2001）来选择簇的数量。其他用在选择簇数目的准则有贝叶斯信息准则（BIC）和 Akiake 信息准则（AIC）。Gap 统计（Tibshirani et al.,2001）是另一种普遍使用用来确定簇数量的方法。它的核心假设是当将数据划分成最佳数量的簇时，最后的划分结果对随机摄动有最大的回弹力。狄利克雷方法（DP）（Ferguson，1973；Rasmussen，2000）引入了簇数量的无参数先验。往往通过概率模型可以导出关于簇数量的后验分布，从这个分布可以计算出最有可能的簇数量。尽管我们有这些目标准则，要找到使得簇最有意义的 K 还是不容易的。图 7a 展示了由 6 种高斯分布成分混合产生的合成数据集。真正的簇标展示在图 7e 中。而这个高斯分布的混合体满足于分别展示在图 7b-d 图中的划分，分别有 2、5、6 种成分，每一种看上去都是合理的。

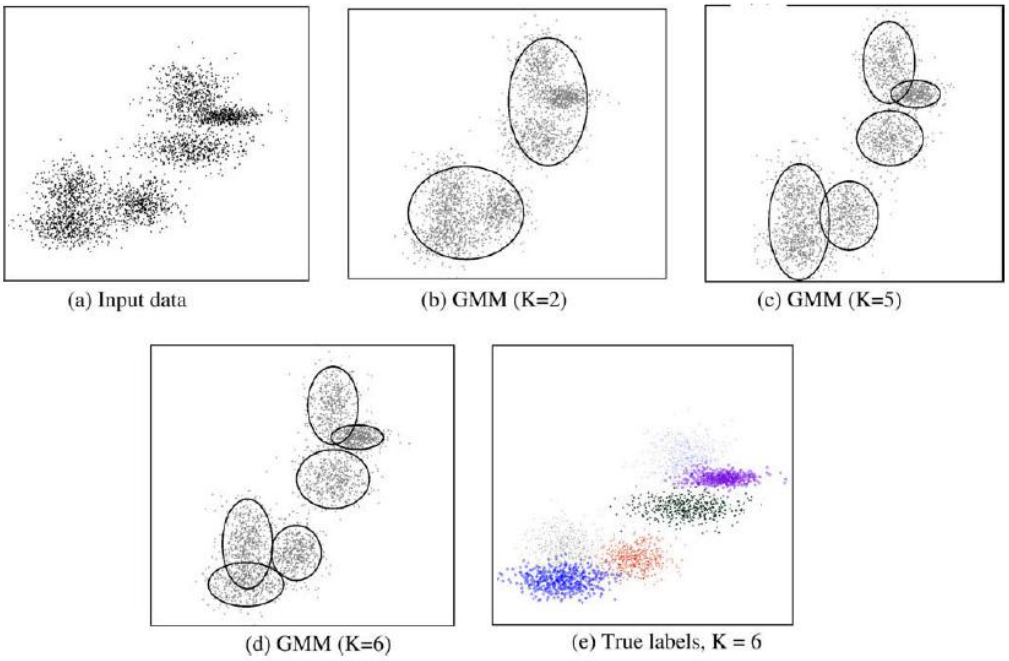

图 7. 簇数目 *K* 的自动选择。图（a）表示由 6 个高斯分布混合产生的输入数据；图（b）-（d）各自表示适用于 2、5、6 个组件的高斯混合模型（GMM）；图（e）表示数据的真实标签。

3.4. 簇的有效性

聚类算法倾向于找到数据中的簇而不考虑是否有簇的存在。图 8a 展示了一个没有自然聚集的数据集；这里所有的点均匀的产生在一个单位正方形中。然而，

图 8b 展示了当 K=3 时，K-means 算法在这些数据上运行时有 3 个簇被识别了出来！簇的有效性指的是以一种定量而且客观的方式（Jain 和 Dubes，1988）评价聚类分析结果的正式程序。事实上，在将聚类算法应用于数据之前，用户就应该确定是否有聚集的趋势（Smith 和 Jain，1984）。

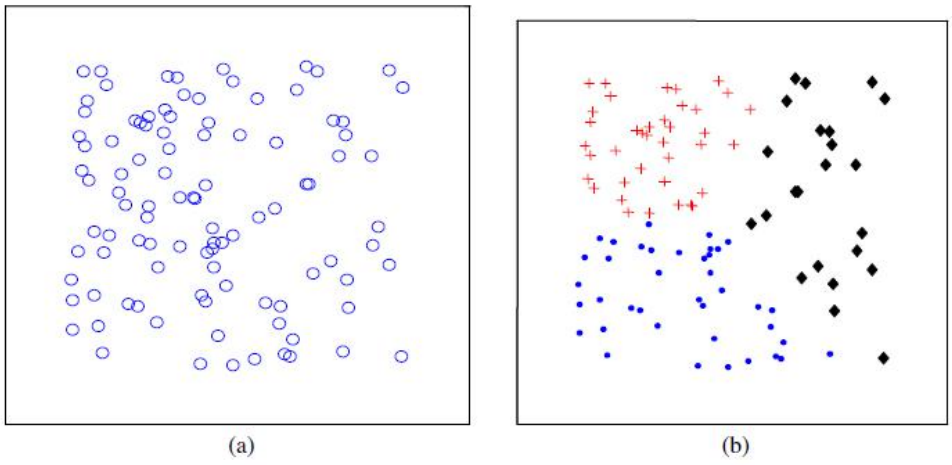

图 8. 聚类有效性。图（a）表示没有自然聚集情况的数据集；图（b）表示 K-means 当 $K = 3$ 时的划分结果。

聚类有效性指标可以基于 3 个不同的准则来定义：内部的、外部的和相对的（Jain 和 Dubes，1988）。基于内部准则的指标仅仅使用数据本身评估聚类算法应用的结构（聚类结果）和数据的合适性。基于相对的指标比较多种结构（比如说由不同的算法产生的结果）然后决定哪个结构在某种程度上更好。基于外部准则的指标通过对照聚类结构和先验信息来衡量，先验信息也就是“真实”的簇标（通常叫做地面真值）。一般来说，聚类结果是用外部准则来评价的，那么如果我们知道真实的簇标，我们又为什么要费功夫去聚类呢？聚类稳定性的概念（Lange et al.2004）非常有吸引力，因为可以用来衡量内部稳定性。聚类稳定性由输入数据在不同二次抽样的情况下聚类结果的差异量来衡量。不同的差异衡量可以获得不同的稳定性衡量。在（Lange et al.,2004）中，通过把从二次抽样聚类得到的簇标作为“真实”簇标，从数据的二次抽样中训练获得有监督的分类器。在测试子集上分类器的性能表明聚类算法的稳定性。在基于模型的算法（比如 K-means 中的基于形心表示法和混合高斯模型）中，从不同的二次抽样中得到的模型之间的距离可以用来衡量稳定性（von Luxburg 和 David，2005）。Shamir 和 Tishby（2008）将稳定性定义为聚类算法的泛化能力（从 PAC 贝叶斯意义上

考虑）。他们认为因为很多算法可以被证明是渐进稳定的，因此关于样本数量的渐进稳定度在衡量聚类稳定性上更有用。交叉检验在评估有监督学习方法时广泛使用。通过用一个不同的有效性衡量的概念替代预测准确性的概念，交叉检验被用在无监督学习上。比如，给定从一个文件的数据中获得的混合模型，其它文件中数据的分布可以作为算法性能的一种显示并被用来决定簇的数目 K。

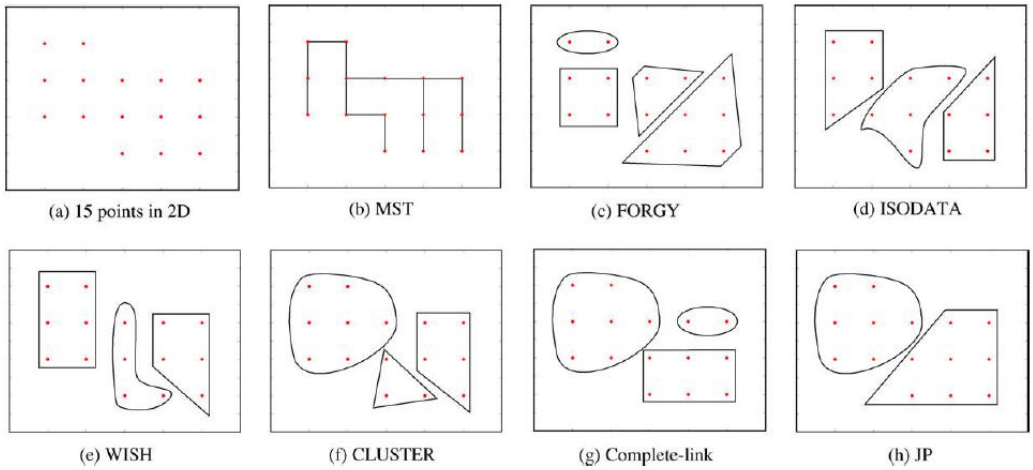

图 9. 15 个二维模式的几个聚类结果：图（a）表示 15 个模式；图（b）表示 15 个模式的最小生成树；图（c）表示 FORGY 的聚类结果；图（d）表示 ISODATA 的聚类结果；图（e）表示 WISH 的聚类结果；图（f）表示 CLUSTER 的聚类结果；图（g）表示完全连接层次聚类的结果；图（h）表示 Jarvis-Patrick 聚类算法的聚类结果。（图取自 Dubes and Jain（1976））

3.5. 比较聚类算法

即使对于相同的数据不同的聚类算法往往会产生完全不同的划分。在图 9 中，7 种不同的算法被用来聚类 15 个二维点，其中 FORGY、ISODATA、CLUSTER 和 WISH 是最小化平方误差准则的基于划分的算法（他们都是基本 K-means 算法的变体）。剩下的 3 种算法，MST（最小生成树）可以看做单连接的基于层次的算法，而 JP 是一种最近邻算法。注意到一个基于层次的算法可以通过确定一个相似度阈值来产生一个划分。很明显，没有哪个聚类结果是优于哪一个的，但是有些却和另一些相似。

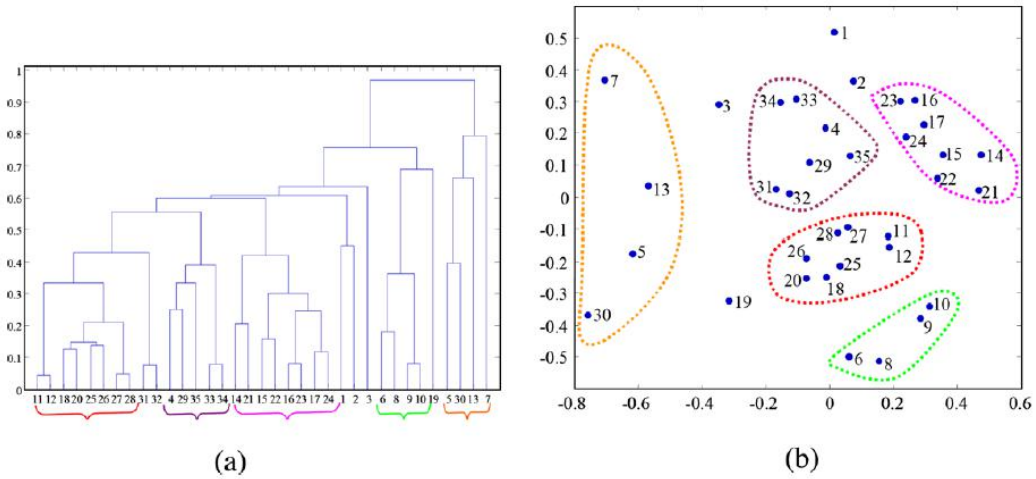

图 10. 聚类算法的聚类结果。图（a）表示 35 种不同算法的层次聚类结果；图（b）表示通过将 35 种算法 Sammon 映射到 2 维空间，其中 的簇被突出用于可视化。其中，组别（4,29，31-35）中的算法是和 K-means，谱聚类，高斯混合模型以及沃德连接算法相关的。组别（6,8-10）中的算法相当于是带有不同目标函数的 CHAMELEON 算法。

一个有趣的问题是如何确定不管数据如何变化，都能产生相似划分的算法。换句话说，我们能对聚类算法进行聚类吗？Jain et al.(2004)基于 35 种不同的算法对 12 个不同数据集的划分将他们聚类成 5 组。一对划分之间的相似度通过使用调整 Rand 指标（ARI）来衡量。35 种聚类算法的一个基于层次的聚类展示在图 10a 中。相关的算法被聚类在一起这并不令人惊讶。为了得到算法间的相似度可视化结果，通过对 35*35 的相似度矩阵应用 Sammon 投影算法（Sammon，1969）得到 35 种算法嵌入在二维空间中的结果。图 10b 展示了所有的 CHAMELEON 变体被聚类到同一个簇中，这幅图表明依据相同聚类策略的聚类算法，尽管在相关参数或者相关目标函数上有小幅差异，却都能产生相似的聚类结果。在（Meila，2003），一个叫做信息变异的术语被提出来作为聚类空间中的另一种尺度。它通过交替聚类时，信息量的丢失或获得来衡量两个聚类算法之间的相似度。

聚类算法也可以基于目标函数在理论层面上进行比较。为了进行这样的比较，聚类方法和聚类算法之间的差异应该被明确（Jain and Dubes，1988）。一个聚类方法是用来解决聚类问题的一个总体策略。而聚类算法仅仅是方法的一个例子。比如，最小化平方误差是一个聚类方法，它涉及很多像 K-means 这样应用该方法的聚类算法。即使在不同的聚类方法之间也可以看到一些等价的关系。比如，

Dhillon et al.（2004）表明谱聚类和核 K-means 是等价的，在选择谱聚类的要点时，其中目标函数的选择和核 K-means 是相同的。（Ding et al.,2005）展示了聚类非负矩阵因子分解和核 K-means 算法的等价性。所有这些方法都和相似度矩阵特征向量分析直接相关。

以上这些讨论说明了聚类的一个很重要的事实：没有最好的聚类算法。每一种聚类算法都或明确或含蓄的给数据一种结构。当模型和数据之间有一个好的匹配时，就能得到一个好的划分。因为数据的结构事先是不知道的，在处理手上的聚类任务时，就需要通过尝试使用竞争的和多样的方法来最终决定合适的算法。这种没有最好的聚类算法的思想和不可能理论部分的相符，表明没有一个聚类算法同时满足数据聚类的所有基本原则。

3.5. 聚类算法的容许性分析

Fisher 和 vanNess（1971）正式以比较聚类算法和为选择聚类步骤提供指导为目标分析了聚类算法。他们定义了一些聚类算法的容许性准则。这些准则测试聚类算法关于不改变本质结构的数据改变的敏感性。如果一个聚类满足准则 A，那么这个聚类叫做 A 容许。有些准则包括凸状点、簇均衡、簇冗余和单调。下面简要描述他们：

(a)凸状：一个聚类算法聚类结果的簇的凸壳并不相交则称该聚类算法是是凸状容许的。

(b)簇均衡：如果聚类算法的结果使一些簇复制任意次数情况下，簇的边界都不改变，那么就称这个聚类算法为簇均衡容许。

(c)簇冗余：如果从聚类算法结果中移除其中一个簇的数据，再运行算法，得到的 K-1 个簇和未再运行时的 K-1 个簇是一样的，那么就称该聚类算法为冗余容许的。

(d)单调：如果当相似度矩阵元素单调改变时，聚类算法结果不发生改变，那么就称该聚类算法是单调容许的。

Fisher 和 Van Ness 证明不能设计出满足一定容许准则的算法。比如，如果一个算法是单调容许的，那么它不能是基于层次的聚类算法。

Kleinberg（2002）指出了一个相似的问题，并定义了三个准则：

(a)标度不变性：任意标度的相似度矩阵不改变聚类结果。

(b)丰富性：聚类算法能获得数据的所有可能划分。

(c)一致性：收缩簇内距离和拉伸簇间距离，聚类结果并不改变。

Kleinberg 也提供了和（Fisher 与 VanNess，1971）相似的结论，表明设计出同时满足所有这些特性的算法是不可能的。因此，他论文的题目是《关于聚类的不可能理论》。（Kleinberg，2002）进一步讨论揭示可以通过将“满足”准则的条件放宽到“近似满足”准则来设计聚类算法。尽管在这里定义的设定在很大程度上是合理的，但他们绝不是唯一可能的设定。因此聚类算法评价的结果必须根据实际情况去解读（Ben-David 和 Ackerman，2008）。

4. 数据聚类的趋势

信息爆炸不仅仅是创造了大量的数据，同时也创造了包括有结构的和无结构的多样化的数据。无结构的数据是指一个并不遵从一定格式的对象集。比如，图像、文本、声频和视频等。另一方面，有结构的数据是指数据对象之间有重要的语义关系。大部分聚类方法忽略用于聚类的对象的结构，并对有结构和无结构的数据采用基于特征向量的表示法。基于向量特征表示法的传统数据划分观点并不能提供充足的框架。比如，当使用点集（Lowe，2004）、消费者购买记录（Guha et al.,2000）、通过问卷和排名获得的数据（Critchlow，1985）、社交网络（Wasserman 和 Faust，1994）以及数据流（Guha et al.,2003b）来代表对象的情形。很多模型和算法被开发用于处理大量的异构数据。数据聚类中最近的一些趋势简要的总结如下：

4.1. 聚类集成

有监督学习的集成方法应用的成功激发了无监督学习集成方法的发展（Fred 和 Jain，2002）。主要思想是，通过对同一数据的多样考察生成同一数据的多样划分（聚类集成）。即使当簇并不是很紧凑而且分离的很好时，也能通过综合划分的结果来获得一个较好的数据划分。Fred 和 Jain 采区这种方式通过 K-means 进行划分的集成。这个集成通过改变不同的 K 值和使用随机的簇初始化得到。

通过使用共生矩阵，这些划分被综合起来产生了一个不错的簇的分离结果。一个聚类集成的例子展示在图 11 中，一个“2—螺旋 ”数据集被用来证明聚类集成的有效性。K-means 在不同的簇数量 K 值下运行了 N 次。一对点之间新的相似度被定义为在 N 次 K-means 运行过程中两个点同时出现在一个簇中的次数。基于新的对间相似度进行聚类获得最终的聚类结果。Strehl 和 Ghosh（2003）提出了集中应用于集成多重划分的概率模型。更新的关于聚类集成的工作可以在（Hore et al.,2009a）中找到。

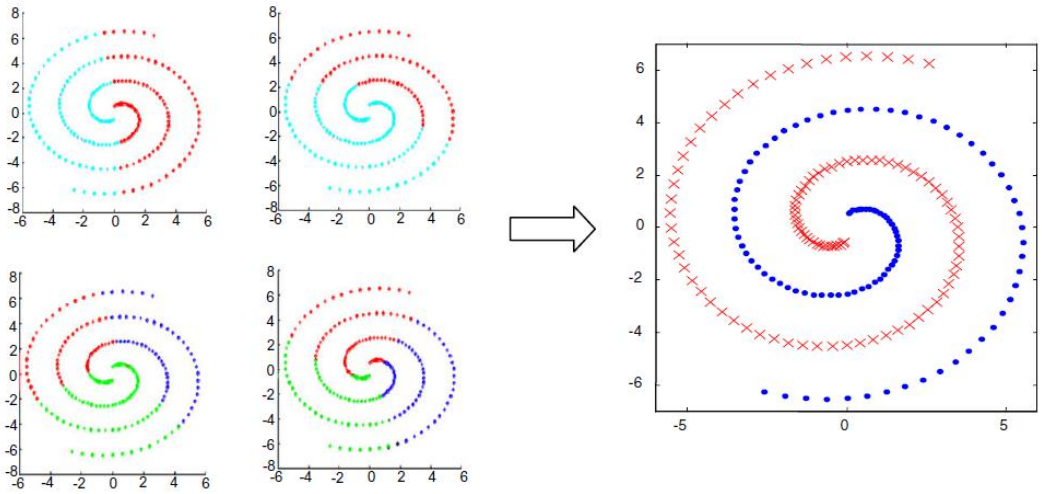

图 11. 聚类集成。通过多次运行 K-means，基于簇中点的“共现”学习对间相似度。这种相似度可以被用来找到任意形状的簇。

有很多方法用于产生聚类集成和综合划分结果。比如多重数据划分可以通过以下方法产生：（i）应用不同的聚类算法。（ii）用同一个聚类算法但是带有不同的参数值或者初始化。（iii）结合不同的数据表示法（特征空间）和聚类算法。结合不同的划分提供的信息，这一证据的积累步骤可以看做是数据点之间相似性测度的学习。

4.2. 半监督聚类

聚类问题本身是一个病态问题。它的目标是只基于固有的信息将数据划分成不知道簇数目的簇。聚类问题数据驱动的特性使得设计出准确找到给定数据的簇变得非常困难。除了 n*d 维模式矩阵或者 n*n 维相似度矩阵之外任何可用的外部或者边带信息都对找到一个好的数据的划分非常有用。使用这样的边带信息的聚

类算法被叫做运行在半监督模式（Chapelle et al.,2006)的聚类算法。这里有两个开放式的问题：（i）应该如何定义边带信息。（ii）在实践中如何获得边带信息。确定边带信息最常见的方法是以对间约束形式。一个“必须连接”约束，确定受约束的某对点属于同一个簇。另一方面，一个“不能连接”约束，确定受约束的某对点不属于同一簇。通常假设这些约束是由领域的专家提供的。从数据自动导出约束的相关工作现在还很有限。有些人尝试通过邻域本体论和其它外部源引出聚类算法的约束包括利用词网本体论、基因本体论和维基百科等来指导聚类求解。然而这些大多是整体特征的约束而不是对于特定实体的约束（Hotho et al.,2003;Liu et al.,2004;Banerjee et al.,2007b）。其它包含边带信息的方法包括（i）“播种”，为了更好的聚类，使用大量无标识的数据中含有一些有标识数据的数据集（Basu et al.,2002）。（ii）允许加强或弱化连接的方法（Law et al.,2005;Figueiredo et al.,2006）。

图 12 是半监督学习在图像分割中应用的一个例子（Lange et al.,2005）。图 12a 表示用于分割（聚类）的纹理图像。除了图像本身，也提供了一系列由用户确定的关于像素点的对间约束。图 12b 是在没有约束使用情况下得到的聚类结果，而图 12c 表示的是在使用约束的情况下聚类结果的改善情况。在两种情况下，簇的数目都假定已知（K=5）。

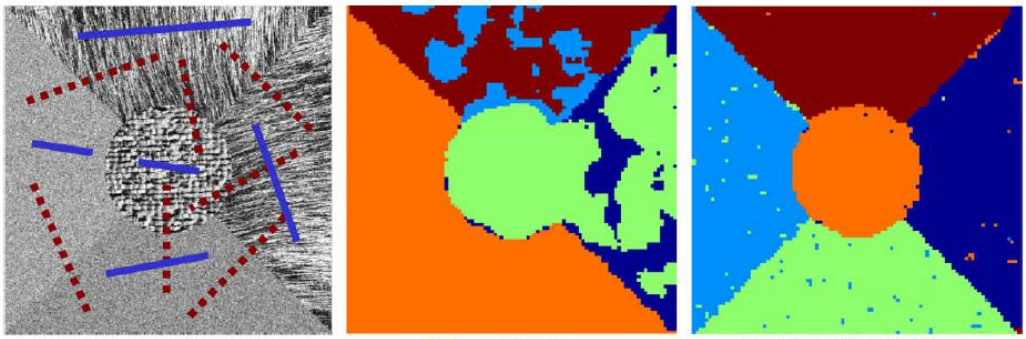

图 12. 半监督学习。图（a）表示由 5 种均匀质地区域组成的输入图像；像素点之间必须连接（蓝色实线）和不能连接（红色断线）约束已被确定。图（b）表示在无约束下的 5 簇解（分割）。图（c）表示在有 10%数据点包含有对间约束时得到的改进的聚类结果（有 5 个簇）。

大部分半监督聚类方法（Bar-Hillel et al.,2003;Basu et al.,2004;Chapelle et

al.,2006;Lu 和 Leen，2007）都通过修正当前聚类算法的目标函数来包含对间约束。我们想要的是能够有一个半监督聚类方法在提高已有聚类算法性能的同时又不改变它。BoostCluster（Liu et al.,2007）采用了这样的观点并遵循一个推动框架，通过使用对间约束去提高任何给定聚类算法的性能。它通过产生新的数据表示法（转换 n*n 的相似度矩阵）迭代的修正聚类算法的输入，这样的话在满足对间约束的同时也保持了聚类输出的完整性。图 13 展示了基于在 UCI 库（Blake，1998）中的有 4000 个 256 维数据的手写数字数据库的 BoostCluster 的性能。BoostCluster 可以提高 3 种用于聚类的常见算法，K-means、单连接、谱分析，当数据带有对间约束时的表现。这里只有“必须连接”约束而且簇的数目假定已知（K=10）。

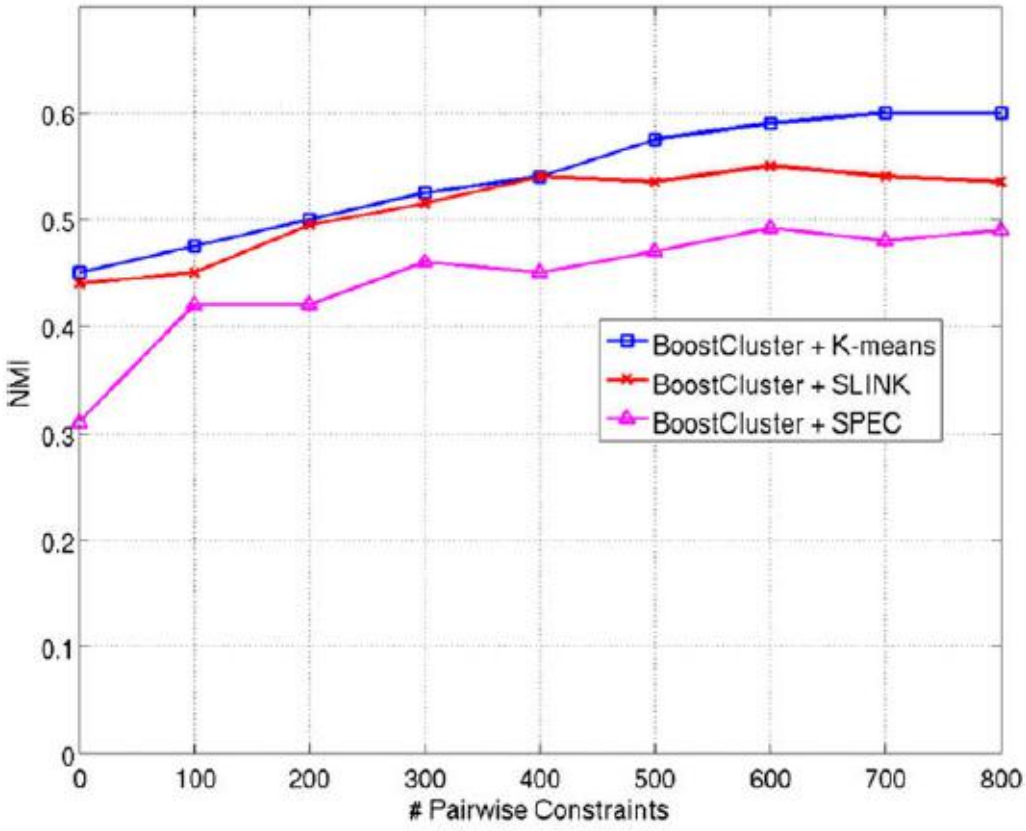

图 13. 随着对间约束增加 BoostCluster（使用标准化的交互信息（NMI）来衡量）的性能变化。三个折现图相当于提高了性能的 K-means、单连接（SLINK）和谱聚类（SPEC）。

4.3. 大规模聚类

大规模数据聚类处理关于由数以千计特征表示的有数以百万计数据点的数据集的聚类问题的挑战。表 1 展示了一些大规模数据聚类实际应用的例子。下面

我们回顾一下大规模数据聚类在基于内容的图像检索中的应用。

表 1. 大规模数据聚类应用的例子

Application	Description	# Objects	# Features
Document clustering	Group documents of similar topics (Andrews et al., 2007)	10^6	10^4
Gene clustering	Group genes with similar expression levels (Lukashin et al., 2003)	10^5	10^2
Content-based image retrieval	Quantize low-level image features (Philbin et al., 2007)	10^9	10^2
Clustering of earth science data	Derive climate indices (Steinbach et al., 2003)	10^5	10^2

基于内容的图像检索（CBIR）的目标是根据给定的查询图像检索出看上去相似的图像。虽然大概过去的 15 年一直在研究这个问题，目前还只有有限的成果。大部分关于 CBIR 的早期工作都是通过计算基于特征的颜色、形状和纹理来定义图像之间的相似度。2008 年一个关于 CBIR 的调查，总结了用于 CBIR 的以时间为序的不同方法（Datta et al.,2008）。最近的用于 CBIR 的方法使用基于特征的关键点。比如，SIFT（Lowe，2004）描述符可以被用来表示图像（见图 14）。然而，一旦图像数据库扩大（大约 1000 万），并假设计算一对图像的匹配时间需要 10ms，一个线性搜索大约需要消耗大学 30h 来响应一个查询指令。这显然是不可接受的。

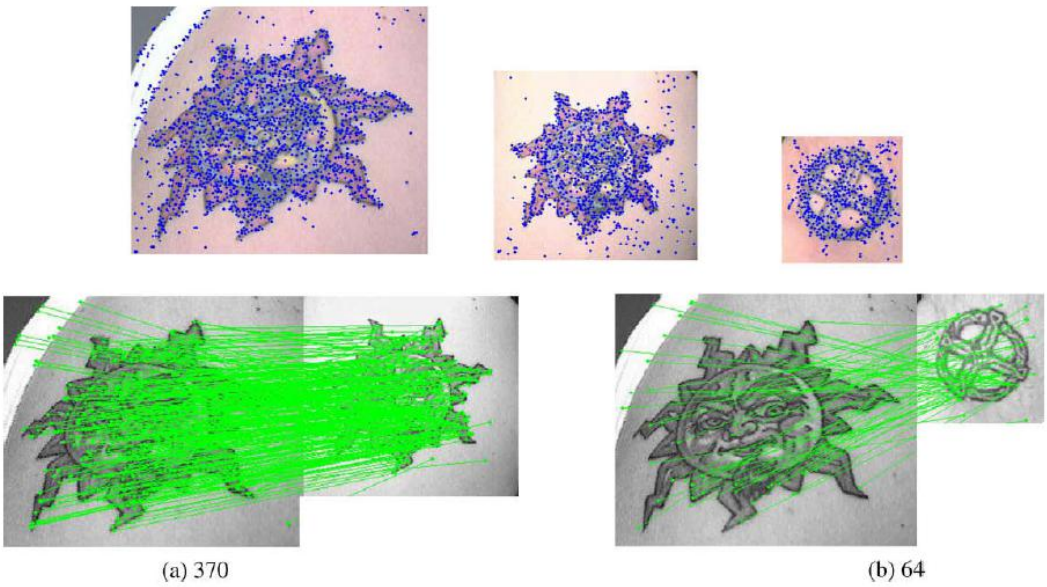

图 14.使用 SIFT 关键点表示的三个纹身。图（a）表示一对相似图像间有 370 个一致的关键点；图（b）表示一对不同图像间有 64 个一致的关键点。绿色的线表示图像间的一致关键点（Lee et al.,2008）。

另一方面，文本检索的应用要快很多。在 Google 大概只需要十分之一秒就能检索 100 亿的文档。一个图像检索的新方法是将问题转化为文本检索问题。所有图像的关键点，首先被聚类成大量的簇（数量上往往比图的关键点少的多）。这些被叫做视觉单词。这样一幅图就可以被视觉单词的柱状图所表示。也就是图的关键点数量就在于每个字或每个簇中。通过用视觉字的柱状图表示每一幅图，我们就将图的搜索问题转化为文本的检索问题，然后开发用于高效图像检索的文本搜索引擎。量化关键点的主要挑战是用于聚类的对象数量。对于有 1000 幅图的图像集，其中平均每幅图有 1000 个关键点，以及 5000 个目标视觉字，需要将 10 万个对象聚类成 5000 个簇。

能够有效处理大规模数据集的聚类算法已经开发了很多了。大部分这些研究可以被分成四类：

(a)有效最近邻（NN）搜索：在任何聚类算法中的一个基本操作是确定每个数据点的簇身份，而这需要 NN 搜索。用于有效 NN 搜索的算法主要是基于树的（比如 kd 树（Moore，1998；Muja 和 Lowe，2009））和基于随机投影的（比如位置敏感散列（Buhler，2001））。

(b)数据概括：它的目标是通过先将大数据集概括为相对小的子集，然后应

用聚类算法对概括后的数据集进行聚类来提高聚类的效率。相关算法包括BIRCH（Zhang et al.,1996）,divide-and-conquer（Steinbach et al.,2000）,核心集K-means（Har-peled 和 Mazumdar，2004），以及粗化方法（Karypis 和 Kumar，1995）。

(c)分布式计算：这一类方法（Dhillon 和 Modha，1999）将数据聚类算法的每个步骤分成许多可以独立计算的程序。这些独立的计算程序将在不同的处理器中平行的执行用以减少整体的计算时间。

(d)增量聚类：这类算法，比如（Bradley et al.,1998）设计对数据进行单次扫描用以提高数据聚类的效率。这和大多数聚类算法形成对比，因为它们在识别数据中心之前需要多次扫描数据点。COBWEB 是一个受欢迎的基于层次的聚类算法，它对可用的数据进行单次扫描并递增的将数据安排到分类树中（Fisher，1987）。

(e)基于抽样的方法：像 CURE（Guha et al.,1998;Kollios et al.,2003）这样的算法,有选择的对大数据集进行二次抽样，并对较小的数据集进行聚类，最后再将结果转化到大的数据集。

4.4. 多方式聚类

被用来聚类的对象或者实体往往是相关的不同成分的一个组合。比如，一个文档是由字、标题、作者和引文等组成。虽然在聚类之前对象的各个成分都可以转化到一个合并的特征向量，但这不是对象的自然表示法，而且很有可能导致糟糕的聚类表现。

共同聚类（Hartigan，1972；Mirkin，1996）试图对数据实体和特征同时进行聚类（或者说对 n*d 维模式矩阵的行和列），也即明确特征子集，有了特征子集，根据一定的评估准则的聚类结果就会有意义。这个问题第一次研究是在Hartigan（1972）写的叫做《直接聚类》中。这也被叫做二维聚类（Cheng et al.,2000）、双聚类、对聚类或者双模聚类。这个概念也和子空间聚类有关，所有的簇在共同的子空间中是确定的。共同聚类在生物信息领域，尤其是基因聚类中最受欢迎，而且也被成功的应用在文本聚类中（Slonim 和 Tishby，2000；Dhillon et al.,2003）。

共同聚类框架被拓展到多方式聚类（Bekkerman et al.,2005），通过聚类对象

集的同时也聚类对象集中的不同成分。这个问题实际上要有挑战性的多，因为不同成分对之间可能有不同的相似度关系。另外，有些关系可能涉及不只两种成分。Banerjee et al.(2007a)提出了一个适用于一类关于被称为 Bregman 分歧的损失函数的多方式聚类方案族。

4.5. 异构数据

在传统的模式识别设定中，一个特征向量由一个对象的不同特性的衡量组成。对几种类型的数据来说这种对象的表示法并不是一种自然表示法。异构数据是指那些不能由固定长度的特征向量自然表示的对象数据。

排名数据：考虑不同的人对 n 部电影排名所生成的数据集，n 个对象中只有一些被排名了。任务是聚类排名相似的用户，而且要明确每一组的代表性排名（Mallows，1957；Critchlow，1985；Busse et al.,2007）。

动态数据：动态数据和静态数据相反，比如博客、网页等会随着时间进程而发生改变。当数据修改之后，聚类结果必须相应的更新。数据流就是一种动态数据，它本质上是瞬态的，不能存储在硬盘上。例子包括路由器接收到的网络包、股票市场、连锁店和信用卡的交易数据流。数据流的特征包括它们的高容量、潜在无界规模、顺序存取以及动态演化。这给传统的聚类算法提出了额外的要求，要求算法能够快速处理和概括大量连续抵达的数据。它还要求算法有适应数据分布改变的能力，侦测新生簇、分辨新生簇和数据中的噪声以及融合旧簇和抛弃失效簇的能力。因为期望是单次扫描算法（Guha et al.,2003b）,上述这些要求就使得数据流聚类成为一个重大的挑战。因为要求高速处理，很多数据流聚类方法（Guha et al.,2003a;Aggarwal et al.,2004;Cao et al.2006;Hore et al.,2009b）都是像 K-means、K-medoid、模糊 c-means 或基于密度聚类的拓展，然后用于数据流环境的设定。

图数据：有些对象，比如说化合物、蛋白质结构等可以非常自然的用图来表示。很多最初的关于图聚类的工作专注于抽象出图的特征，从而可以使用已有的聚类算法来处理图的特征向量（Tsuda 和 Kudo，2006）。特征的提取可以是像频繁子图、最短路径、循环和基于树这样基于模式的。伴随着核学习的兴起，有越来越多的工作专注于更适合基于图的数据的核函数的开发。一种确定图之间相

似度的方法是通过调整相应邻接矩阵的表示法（Umeyama，1988）。

关系型数据：另一个吸引大量兴趣的方面是聚类关系型（网络）数据。不像以划分图集为不相交的组为目的的聚类图数据，这里的任务是划分一个大图（比如网络）为基于它们连接结构和节点属性的紧密结合的子图。这个问题当允许连接（对象之间的关系）为不同的类型时将变得更加复杂。对于关系型数据聚类来说，一个核心问题是要定义一个合适的聚类准则。（Taskar et al.,2001）第一次提出了关于关系型数据的通用概率模型，它依据受相互之间约束的分布对不同的相关实体进行建模。纽曼模块性函数（Newman 和 Girvan，2004；Newman，2006）是一个广泛用于寻找网络间群体结构的准则，但是它只考虑了连接结构而忽略了属性相似度。（White 和 Smyth，2005）提出了用于网络图聚类的纽曼格文目标函数（Newman 和 Girvan，2004）的谱松弛。因为实际的网络往往是动态的，另一个问题是要在考虑组成员关系和其它特征特性改变的情况下对网络演化行为的建模（Backstrom et al.,2006）。

5. 总结

对数据进行合理的分组问题很自然的出现在很多科学领域。因此见证数据聚类的不断流行并不令人惊讶。记住聚类分析是一种探索性的工具将非常重要。聚类算法的输出仅仅是一种建议性的假设。尽管大量的聚类算法已经出现或者正在出现，没有某一个聚类算法已经被证明在不同应用领域中优于其它算法。很多包括简单 K-means 在内的算法都是容许性算法。随着新应用的不断出现，人们渐渐的意识到寻找最好的聚类算法这个问题本身就是有问题的。比如，考虑企业知识管理这样一个应用领域。给定同一文档库，不同的用户群（比如法律、营销、管理等）也许就只对基于各自需要的文档划分有兴趣。一个满足一群用户需要的聚类方法也许就不满足另一群用户的需要。就像前面提到的“聚类结果在于观察者眼中”，所以实际数据聚类必须结合用户和应用需要。

聚类在数据分析领域有许多成功的案例。尽管如此，机器学习和模式识别领域还需要处理许多问题，用于提高我们对数据聚类的理解。从这一点而言，以下是一系列有价值的问题和研究方向。

(a)研究群体需要可用的一系列基准数据（有地面事实的）来测试评估聚类方

法。基准数据应该来自不同的领域（文档、图像、时间序列、消费者交易记录、生物序列、社交网络等）。基准数据还应该包括静态数据和动态数据（后者将对分析随时间变化的簇非常有用）、定性的和定量的属性以及连接和非连接的对象等。尽管提供基准数据的想法并不新鲜（比如 UCI ML 和 KDD repository），但目前基准数据还主要是小的、静态的数据集。

(b)我们需要聚类算法和应用需要之间更加紧密的整合。比如，有些应用只需要产生一些紧密结合的簇（并不是紧密结合的簇可以被忽略），而有些应用需要获得对整个数据集进行最好的划分。在大部分应用中，并不一定要找到最佳的聚类算法，更重要的是选择用于识别数据潜在簇结构的正确特征提取方法。

(c)无论原则（或目标），大部分聚类方法最后都转化为组合优化问题，它的目的是找到优化目标的数据划分。因此，在涉及大规模数据时，计算问题就变得非常重要。比如，找到 K-means 的全局最优解就是一个 NP 困难问题。因此选择能够使得计算上有效率求解的聚类原则就变得非常重要。

(d)关于聚类的一个基本问题是它的稳定性和一致性。一个好的聚类原则应该使得在数据中有扰动情况下产生的数据划分是稳定的。我们需要开发出能够产生稳定解的聚类方法。

(e)根据对已有公设的满足来选择聚类原则。尽管有 Kleinberg 的不可能理论，有些研究已经表明可以通过放松一些设定来满足要求。因此，也许评估一种聚类原则的方式就是看它满足设定的程度。

(f)考虑到聚类本身的困难，这使得开发半监督聚类算法变得更有意义。它可以使用有标识的数据和（用户确定的）对间约束来决定（i）数据表示法（ii）用于数据聚类的合适的目标函数。

致谢

我要感谢国家科学基金和海军研究办公室对我关于数据聚类、降维、分类和半监督学习的研究工作的支持。我要感谢 Rong Jin、Pang-Ning Tan 和 Pavan Mallapragada 帮助我准备傅京孙演讲和这个手稿。我很享受和 Eric Backer 等在数据聚类上富有成效的合作，并且学到了很多。Joydeep Ghosh 等为提高本文的的质量提供了很多有用的建议。

参考文献

[1] Aggarwal, Charu C., Han, Jiawei, Wang, Jianyong, Yu, Philip S., 2003. A framework for clustering evolving data streams. In: Proc. 29th Internat. Conf. on Very Large Data Bases, pp. 81–92.

[2] Agrawal, Rakesh, Gehrke, Johannes, Gunopulos, Dimitrios, Raghavan, Prabhakar, 1998. Automatic subspace clustering of high dimensional data for data mining applications. In: Proc. ACM SIGMOD, pp. 94–105.

[3] Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press. Andrews, Nicholas, O., Fox, Edward A., 2007. Recent developments in documentclustering. Technical report TR-07-35. Deparment of Computer Science, Virginia Tech.

[4]Arabie, P., Hubert, L., 1994. Cluster analysis in marketing research. In: Advanced Methods in Marketing Research. Blackwell, Oxford, pp. 160–189.

[5]Backer, Eric, 1978. Cluster Analysis by Optimal Decomposition of Induced Fuzzy Sets. Delft University Press.

[6]Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X., 2006. Group formation in large social networks: Membership, growth, and evolution. In: Proc. 12th KDD.

[7]Baldi, P., Hatfield, G., 2002. DNA Microarrays and Gene Expression. Cambridge University Press.

[8]Ball, G., Hall, D., 1965. ISODATA, a novel method of data anlysis and pattern classification. Technical report NTIS AD 699616. Stanford Research Institute, Stanford, CA.

[9]Banerjee, Arindam, Merugu, Srujana, Dhillon, Inderjit, Ghosh, Joydeep. 2004. Clustering with bregman divergences. J. Machine Learn. Res., 234–245.

[10]Banerjee, Arindam, Basu, Sugato, Merugu, Srujana, 2007a. Multi-way clustering on relation graphs. In: Proc. 7th SIAM Internat. Conf. on Data Mining.

[11]Banerjee, S., Ramanathan, K., Gupta, A., 2007b. Clustering short texts using Wikipedia. In: Proc. SIGIR.

[12]Bar-Hillel, Aaron, Hertz, T., Shental, Noam, Weinshall, Daphna, 2003. Learning distance functions using equivalence relations. In: Proc. 20th Internat. Conf. on Machine Learning, pp. 11–18.

[13]Basu, Sugato, Banerjee, Arindam, Mooney, Raymond, 2002. Semi-supervised clustering by seeding. In: Proc. 19th Internat. Conf. on Machine Learning.

[14]Basu, Sugato, Bilenko, Mikhail, Mooney, Raymond J., 2004. A probabilistic framework for semi-supervised clustering. In: Proc. 10th KDD, pp. 59–68.

[15]Basu, Sugato, Davidson, Ian, Wagstaff, Kiri (Eds.), 2008. Constrained Clustering: Advances in Algorithms, Theory and Applications. Data Mining and Knowledge Discovery, vol. 3, Chapman & Hall/CRC.

[16]Bekkerman, Ron, El-Yaniv, Ran, McCallum, Andrew, 2005. Multi-way distributional clustering via pairwise interactions. In: Proc. 22nd Internat. Conf. Machine Learning, pp. 41–48.

[17]Belkin, Mikhail, Niyogi, Partha, 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing Systems, vol. 14. pp. 585–591.

[18]Ben-David, S., Ackerman, M., 2008. Measures of clustering quality: A working set of axioms for clustering. Advances in Neural Information Processing Systems.

[19]Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press.

[20]Bhatia, S., Deogun, J., 1998. Conceptual clustering in information retrieval. IEEE Trans. Systems Man Cybernet. 28 (B), 427–436.

[21]Bishop, Christopher M., 2006. Pattern Recognition and Machine Learning. Springer. Blake, Merz C.J., 1998. UCI repository of machine learning databases.

[22]Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Machine Learn. Res. 3, 993–1022.

[23]Bradley, P.S., Fayyad, U., Reina, C., 1998. Scaling clustering algorithms to large databases. In: Proc. 4th KDD.

[24]Buhler, J., 2001. Efficient large-scale sequence comparison by locality-sensitive hashing. Bioinformatics 17 (5), 419–428.

[25]Busse, Ludwig M., Orbanz, Peter, Buhmann, Joachim M., 2007. Cluster analysis of heterogeneous rank data. In: Proc. 24th Internat. Conf. on Machine Learning, pp. 113–120.

[26]Cao, F., Ester, M., Qian, W., Zhou, A., 2006. Density-based clustering over an evolving data stream with noise. In: Proc. SIAM Conf. Data Mining.

[27]Chapelle, O., Schölkopf, B., Zien, A. (Eds.), 2006. Semi-Supervised Learning. MIT Press, Cambridge, MA.

[28]Cheng, Yizong, Church, George M., 2000. Biclustering of expression data. In: Proc. Eighth Internat. Conf. on Intelligent Systems for Molecular Biology, AAAI Press, pp. 93–103.

[29]Connell, S.D., Jain, A.K., 2002. Writer adaptation for online handwriting recognition. IEEE Trans. Pattern Anal. Machine Intell. 24 (3), 329–346.

[30]Critchlow, D., 1985. Metric Methods for Analyzing Partially Ranked Data. Springer. Datta, Ritendra, Joshi, Dhiraj, Li, Jia, Wang, James Z., 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40 (2) (Article5).

[31]Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. 39, 1–38.

[32]Dhillon, I., Modha, D., 1999. A data-clustering algorithm on distributed memory multiprocessors. In: Proc. KDD'99 Workshop on High Performance Knowledge Discovery, pp. 245–260.

[33]Dhillon, Inderjit S., Mallela, Subramanyam, Guyon, Isabelle, Elisseeff, André, 2003. A divisive information-theoretic feature clustering algorithm for text classification. J. Machine Learn. Res. 3, 2003.

[34]Dhillon, Inderjit S., Guan, Yuqiang, Kulis, Brian, 2004. Kernel k-means: Spectral clustering and normalized cuts. In: Proc. 10th KDD, pp. 551–556.

[35]Ding, Chris, He, Xiaofeng, Simon, Horst D., 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proc. SIAM Internat.

Conf. on Data Mining, pp. 606–610.

[36]Drineas, P., Frieze, A., Kannan, R., Vempala, S., Vinay, V., 1999. Clustering large graphs via the singular value decomposition. Machine Learn. 56 (1–3), 9–33.

[37]Dubes, Richard C., Jain, Anil K., 1976. Clustering techniques: User's dilemma. Pattern Recognition, 247–260.

[38]Duda, R., Hart, P., Stork, D., 2001. Pattern Classification, second ed. John Wiley and Sons, New York.

[39]Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybernet. 3, 32–57.

[40]Eschrich, S., Ke, Jingwei, Hall, L.O., Goldgof, D.B., 2003. Fast accurate fuzzy clustering through data reduction. IEEE Trans. Fuzzy Systems 11 (2), 262–270.

[41]Ester, Martin, Peter Kriegel, Hans, S., Jörg, Xu, Xiaowei, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. 2nd KDD, AAAI Press.

[42]Ferguson, Thomas S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Statist. 1, 209–230.

[43]Figueiredo, Mario, Jain, Anil K., 2002. Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Machine Intell. 24 (3), 381–396.

[44]Figueiredo, M.A.T., Chang, D.S., Murino, V., 2006. Clustering under prior knowledge with application to image segmentation. Adv. Neural Inform. Process. Systems 19, 401–408.

[45]Fisher, Douglas H., 1987. Knowledge acquisition via incremental conceptual clustering. Machine Learn., 139–172.

[46]Fisher, L., vanNess, J., 1971. Admissible clustering procedures. Biometrika.

Forgy, E.W., 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. Biometrics 21, 768–769.

[47]Frank, Ildiko E., Todeschini, Roberto, 1994. Data Analysis Handbook. Elsevier Science Inc., pp. 227–228.

[48]Fred, A., Jain, A.K., 2002. Data clustering using evidence accumulation. In: Proc. Internat. Conf. Pattern Recognition (ICPR).

[49]Frigui, H., Krishnapuram, R., 1999. A robust competitive clustering algorithm with applications in computer vision. IEEE Trans. Pattern Anal. Machine Intell. 21, 450–465.

[50]Gantz, John F., 2008. The diverse and exploding digital universe. Available onlineat:<http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>.

[51]Google Scholar, 2009 (February). Google Scholar. <http://scholar.google.com>. [52]Guha, Sudipto, Rastogi, Rajeev, Shim, Kyuseok, 1998. CURE: An efficient clustering algorithm for large databases. In: Proc. ICDM., pp. 73–84.

[53]Guha, Sudipto, Rastogi, Rajeev, Shim, Kyuseok, 2000. Rock: A robust clustering algorithm for categorical attributes. Inform. Systems 25 (5), 345–366.

[54]Guha, Sudipto, Meyerson, A., Mishra, Nina, Motwani, Rajeev, O'Callaghan, L., 2003a. Clustering data streams: Theory and practice. Trans. Knowledge Discovery Eng.

[55]Guha, Sudipto, Mishra, Nina, Motwani, Rajeev, 2003b. Clustering data streams. IEEE Trans. Knowledge Data Eng. 15 (3), 515–528.

[56]Hagen, L., Kahng, A.B., 1992. New spectral methods for ratio cut partitioning and clustering. IEEE Trans. Comput.-Aid. Des. Integrated Circuits Systems 11 (9), 1074–1085.

[57]Han, Jiawei, Kamber, Micheline, 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.

[58]Hansen, Mark H., Yu, Bin, 2001. Model selection and the principle of minimum description length. J. Amer. Statist. Assoc. 96 (454), 746–774.

[59]Har-peled, Sariel, Mazumdar, Soham, 2004. Coresets for k-means and k-median clustering and their applications. In: Proc. 36th Annu. ACM Sympos. Theory Comput., pp. 291–300.

[60]Hartigan, J.A., 1972. Direct clustering of a data matrix. J. Amer. Statist. Assoc. 67 (337), 123–132.

[61]Hartigan, J.A., 1975. Clustering Algorithms. John Wiley and Sons.Hofmann, T., Buhmann, J.M., 1997. Pairwise data clustering by deterministic annealing. IEEE

Trans. Pattern Anal. Machine Intell. 19 (1), 1–14.

[62]Hore, Prodip, Hall, Lawrence O., Goldgof, Dmitry B., 2009a. A scalable framework for cluster ensembles. Pattern Recognition 42 (5), 676–688.

[63]Hore, Prodip, Hall, Lawrence O., Goldgof, Dmitry B., Gu, Yuhua, Maudsley, Andrew A., Darkazanli, Ammar, 2009b. A scalable framework for segmenting magnetic resonance images. J. Signal Process. Systems 54 (1–3), 183–203.

[64]Hotho, A., Staab, S., Stumme, G., 2003. Ontologies to improve text document clustering. In: Proc. of the ICDM.

[65]Hu, J., Ray, B.K., Singh, M., 2007. Statistical methods for automated generation of service engagement staffing plans. IBM J. Res. Dev. 51 (3), 281–293.

[66]Iwayama, M., Tokunaga, T., 1995. Cluster-based text categorization: A comparison of category search strategies. In: Proc. 18th ACM Internat. Conf. on Research and Development in Information Retrieval, pp. 273–281.

[67]Jain, Anil K., Dubes, Richard C., 1988. Algorithms for Clustering Data. Prentice Hall. Jain, Anil K., Flynn, P., 1996. Image segmentation using clustering. In: Advances in Image Understanding. IEEE Computer Society Press, pp. 65–83.

[68]Jain, A.K., Topchy, A., Law, M.H.C., Buhmann, J.M., 2004. Landscape of clustering algorithms. In: Proc. Internat. Conf. on Pattern Recognition, vol. 1, pp. 260–263.

[69]JSTOR, 2009. JSTOR. <http://www.jstor.org>.

[70]Karypis, George, Kumar, Vipin, 1995. A fast and high quality multilevel scheme for partitioning irregular graphs. In: Proc. Internat. Conf. on Parallel Processing, pp. 113–122.

[71]Kashima, H., Tsuda, K., Inokuchi, A., 2003. Marginalized Kernels between labeled graphs. In: Proc. 20th Internat. Conf. on Machine Learning, pp. 321–328.

[72]Kashima, H., Hu, J., Ray, B., Singh, M., 2008. K-means clustering of proportional data using L1 distance. In: Proc. Internat. Conf. on Pattern Recognition, pp. 1–4.

[73]Kaufman, Leonard, Rousseeuw, Peter J., 2005. Finding groups in data: An introduction to cluster analysis. Wiley series in Probability and Statistics.Kleinberg, Jon, 2002. An impossibility theorem for clustering. In: NIPS

15. pp. 463– 470.

[74]Kollios, G., Gunopulos, D., Koudas, N., Berchtold, S., 2003. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans. Knowledge Data Eng. 15 (5), 1170–1187.

[75]Lange, Tilman, Roth, Volker, Braun, Mikio L., Buhmann, Joachim M., 2004. Stability-based validation of clustering solutions. Neural Comput. 16 (6), 1299–1323.

[76]Lange, T., Law, M.H., Jain, A.K., Buhmann, J., 2005. Learning with constrained and unlabelled data. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition 1, 730–737.

[77]Law, Martin, Topchy, Alexander, Jain, A.K., 2005. Model-based clustering with probabilistic constraints.In: Proc. SIAM Conf. on Data Mining, pp. 641–645.

[78]Lee, Jung-Eun, Jain, Anil K., Jin, Rong, 2008. Scars, marks and tattoos (SMT): Soft biometric for suspect and victim identification. In: Proceedings of the Biometric Symposium.Li, W., McCallum, A., 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proc. 23rd Internat. Conf. on Machine Learning, pp. 577– 584.

[79]Linde, Y., Buzo, A., Gray, R., 1980. An algorithm for vector quantizer design. IEEE Trans. Comm. 28, 84–94.

[80]Liu, J., Wang, W., Yang, J., 2004. A framework for ontology-driven subspace clustering. In: Proc. KDD.

[81]Liu, Yi, Jin, Rong, Jain, A.K., 2007. Boostcluster: Boosting clustering by pairwise constraints. In: Proc. 13th KDD, pp. 450–459.

[82]Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inform. Theory 28, 129–137.

[83]Originally as an unpublished Bell laboratories Technical Note (1957).Lowe, David G., 2004. Distinctive image features from scale-invariant keypoints. Internat. J. Comput. Vision 60 (2), 91–110.

[84]Lu, Zhengdong, Leen, Todd K., 2007. Penalized probabilistic clustering. Neural Comput. 19 (6), 1528–1567.

[85]Lukashin, A.V., Lukashev, M.E., Fuchs, R., 2003. Topology of gene expression networks as revealed by data mining and modeling. Bioinformatics 19 (15), 1909–1916.

[86]MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Fifth Berkeley Symposium on Mathematics. Statistics and Probability. University of California Press, pp. 281–297.

[87]Mallows, C.L., 1957. Non-null ranking models. Biometricka 44, 114–130.

[88]Mao, J., Jain, A.K., 1996. A self-organizing network for hyper-ellipsoidal clustering (HEC). IEEE Trans. Neural Networks 7 (January), 16–29.

[89]McLachlan, G.L., Basford, K.E., 1987. Mixture Models: Inference and Applications to Clustering. Marcel Dekker.

[90]Meila, Marina, 2003. Comparing clusterings by the variation of information. In: COLT, pp. 173–187.

[91]Meila, Marina, 2006. The uniqueness of a good optimum for k-means. In: Proc. 23rd Internat. Conf. Machine Learning, pp. 625–632.

[92]Meila, Marina, Shi, Jianbo, 2001. A random walks view of spectral segmentation. In: Proc. AISTATAS.

[93]Merriam-Webster Online Dictionary, 2008. Cluster analysis. <http://www.merriam-webster-online.com>.

[94]Mirkin, Boris, 1996. Mathematical Classification and Clustering. Kluwer Academic Publishers.

[95]Moore, Andrew W., 1998. Very fast EM-based mixture model clustering using multiresolution kd-trees. In: NIPS, pp. 543–549.

[96]Motzkin, T.S., Straus, E.G., 1965. Maxima for graphs and a new proof of a theorem of Turan. Canadian J. Math. 17, 533–540.

[97]Muja, M., Lowe, D.G., 2009. Fast approximate nearest neighbors with automatic algorithm configuration. In: Proc. Internat. Conf. on Computer Vision Theory and Applications (VISAPP'09).

[98]Newman, M.E.J., 2006. Modularity and community structure in networks. In: Proc. National Academy of Sciences, USA.

[99]Newman, M., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69 (026113), 3.

[100]Ng, Andrew Y., Jordan, Michael I., Weiss, Yair, 2001. On spectral clustering: Analysis and an algorithm. Adv. Neural Inform. Process. Systems, vol. 14. MIT Press, pp. 849–856.

[101]Pampalk, Elias, Dixon, Simon, Widmer, Gerhard, 2003. On the evaluation of perceptual similarity measures for music. In: Proc. Sixth Internat. Conf. on Digital Audio Effects (DAFx-03). pp. 7–12.

[102]Pavan, Massimilliano, Pelillo, Marcello, 2007. Dominant sets and pairwise clustering. IEEE Trans. Pattern Anal. Machine Intell. 29 (1), 167–172.

[103]Pelleg Dan, Moore Andrew, 1999. Accelerating exact k-means algorithms with geometric reasoning. In: Chaudhuri Surajit, Madigan David (Eds.), Proc. Fifth Internat. Conf. on Knowledge Discovery in Databases, AAAI Press, pp. 277–281.

[104]Pelleg, Dan, Moore, Andrew, 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In: Proc. Seventeenth Internat. Conf. on Machine Learning. pp. 727–734.

[105]Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2007. Object retrieval with large vocabularies and fast spatial matching. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition.

[106]Rasmussen, Carl, 2000. The infinite gaussian mixture model. Adv. Neural Inform. Process. Systems 12, 554–560.

[107]Roberts Stephen J., Holmes, Christopher, Denison, Dave, 2001. Minimum-entropy data clustering using reversible jump Markov chain Monte Carlo. In: Proc. Internat. Conf. Artificial Neural Networks. pp. 103–110.

[108]Sahami, Mehran, 1998. Using Machine Learning to Improve Information Access. Ph.D. Thesis, Computer Science Department, Stanford University.

Sammon Jr., J.W., 1969. A nonlinear mapping for data structure analysis. IEEE Trans. Comput. 18, 401–409.

[109]Scholkopf, Bernhard, Smola, Alexander, Muller, Klaus-Robert, 1998. Nonlinear

component analysis as a kernel eigenvalue problem. Neural Comput. 10 (5), 1299–1319.

[110]Shamir, Ohad, Tishby, Naftali, 2008. Cluster stability for finite samples. Adv. Neural Inform. Process. Systems 20, 1297–1304.

[111]Shi, Jianbo, Malik, Jitendra, 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Machine Intell. 22, 888–905.

[112]Sindhwani, V., Hu, J., Mojsilovic, A., 2008. Regularized co-clustering with dual supervision. In: Advances in Neural Information Processing Systems.

[113]Slonim, Noam, Tishby, Naftali, 2000. Document clustering using word clusters via the information bottleneck method. In: ACM SIGIR 2000, pp. 208–215.

[114]Smith, Stephen P., Jain, Anil K., 1984. Testing for uniformity in multidimensional data. IEEE Trans. Pattern Anal. Machine Intell. 6 (1), 73–81.

[115]Sokal, Robert R., Sneath, Peter H.A., 1963. Principles of Numerical Taxonomy. W.H. Freeman, San Francisco.

[116]Steinbach, M., Karypis, G., Kumar, V., 2000. A comparison of document clustering techniques. In: KDD Workshop on Text Mining.

[117]Steinbach, Michael, Tan, Pang-Ning, Kumar, Vipin, Klooster, Steve, Potter, Christopher, 2003. Discovery of climate indices using clustering. In: Proc. Ninth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining.

[118]Steinhaus, H., 1956. Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci. IV (C1.III), 801–804.

[119]Strehl, Alexander, Ghosh, Joydeep, 2003. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. J. Machine Learn. Res. 3, 583–617.

[120]Tabachnick, B.G., Fidell, L.S., 2007. Using Multivariate Statistics, fifth ed. Allyn and Bacon, Boston.

[121]Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin, 2005. Introduction to Data Mining, first ed. Addison-Wesley Longman Publishing Co. Inc., Boston, MA, USA.

Taskar, B., Segal, E., Koller, D., 2001.

[122]Probabilistic clustering in relational data. In: Proc. Seventeenth Internat. Joint Conf. on Artificial Intelligence (IJCAI), pp. 870– 887.

[123]Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. Roy. Statist. Soc. B, 411–423.

[124]Tishby, Naftali, Pereira, Fernando C., Bialek, William, 1999. The information bottleneck method. In: Proc. 37th Allerton Conf. on Communication, Control and Computing, pp. 368–377.

[125]Tsuda, Koji, Kudo, Taku, 2006. Clustering graphs by weighted substructure mining. In: Proc. 23rd Internat. Conf. on Machine Learning. pp. 953–960.

[126]Tukey, John Wilder, 1977. Exploratory Data Analysis. Addison-Wesley.

[127]Umeyama, S., 1988. An eigen decomposition approach to weighted graph matching problems. IEEE Trans. Pattern Anal. Machine Intell. 10 (5), 695–703.

[128]von Luxburg, U., David, Ben S., 2005. Towards a statistical theory of clustering. In: Pascal Workshop on Statistics and Optimization of Clustering.

[129]Wallace, C.S., Boulton, D.M., 1968. An information measure for classification. Comput. J. 11, 185–195.

[130]Wallace, C.S., Freeman, P.R., 1987. Estimation and inference by compact coding (with discussions). JRSSB 49, 240–251.

[131]Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press.

[132]Welling, M., Rosen-Zvi, M., Hinton, G., 2005. Exponential family harmoniums with an application to information retrieval. Adv. Neural Inform. Process. Systems 17, 1481–1488.

[133]White, Scott, Smyth, Padhraic, 2005. A spectral clustering approach to finding communities in graph. In: Proc. SIAM Data Mining.

[134]Yu, Stella X., Shi, Jianbo, 2003. Multiclass spectral clustering. In: Proc. Internat. Conf. on Computer Vision, pp. 313–319.

[135]Zhang, Tian, Ramakrishnan, Raghu, Livny, Miron. 1996. BIRCH: An efficient data clustering method for very large databases. In: Proc.1996 ACM SIGMOD Internat. Conf. on Management of data, vol. 25, pp. 103–114.