# 聚类分析

维基百科，自由的百科全书

**聚类分析**（英语：Cluster analysis，亦称为**群集分析**）是对于统计数据分析的一门技术，在许多领域受到广泛应用，包括机器学习，数据挖掘，模式识别，图像分析以及生物信息。聚类是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集（subset），这样让在同一个子集中的成员对象都有相似的一些属性，常见的包括在坐标系中更加短的空间距离等。

一般把数据聚类归纳为一种非监督式學習。

## 目录



## 聚类类型

数据聚类算法可以分为结构性或者分散性。结构性算法利用以前成功使用过的聚类器进行分类，而分散型算法则是一次确定所有分类。结构性算法可以**从上至下**或者**从下至上**双向进行计算。**从下至上**算法从每个对象作为单独分类开始，不断融合其中相近的对象。而**从上至下**算法则是把所有对象作为一个整体分类，然后逐渐分小。

分散式聚类算法，是一次性确定要产生的类别，这种算法也已应用于**从下至上**聚类算法。

基于密度的聚类算法，是为了挖掘有任意形状特性的类别而发明的。此算法把一个类别视为数据集中大于某阈值的一个区域。DBSCAN和OPTICS是两个典型的算法。

许多聚类算法在执行之前，需要指定从输入数据集中产生的分类个数。除非事先准备好一个合适的值，否则必须决定一个大概值，关于这个问题已经有一些现成的技术。

# 距离测量

在结构性聚类中，关键性的一步就是要选择测量的距离。一个简单的测量就是使用曼哈顿距离，它相当于每个变量的绝对差值之和。该名字的由来起源于在纽约市区测量街道之间的距离就是由人步行的步数来确定的。

一个更为常见的测量是欧式空间距离，他的算法是找到一个空间，来计算每个空间中点到原点的距离，然后对所有距离进行换算。

常用的几个距离计算方法：

- 欧式距离（2-norm距离）
- 曼哈顿距离（Manhattan distance, 1-norm距离）
- infinity norm
- 马氏距离
- 余弦相似性
- 汉明距离

# 结构性聚类

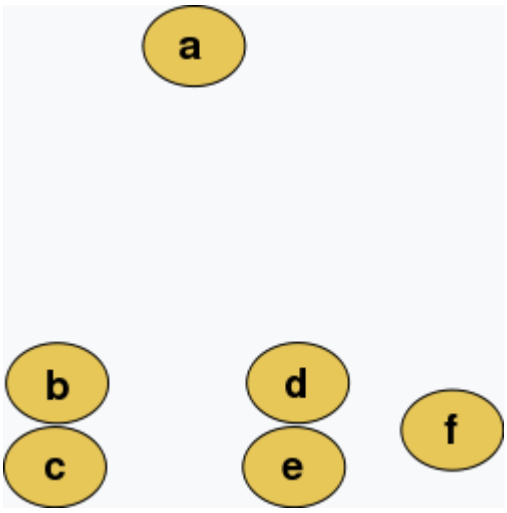
在已经得到距离值之后，元素间可以被联系起来。通过分离和融合可以构建一个结构。传统上，表示的方法是树形数据结构， 然后对该结构进行修剪。树的根节点表示一个包含所有项目的类别，树叶表示与个别的项目相关的类别。

层次聚类算法，要么是自底向上聚集型的，即从叶子节点开始，最终汇聚到根节点；要么是自顶向下分裂型的，即从根节点开始，递归的向下分裂。

任意非负值的函数都可以用于衡量一对观测值之间的相似度。决定一个类别是否分裂或者合并的是一个连动的标准，它是两两观测值之间距离的函数。

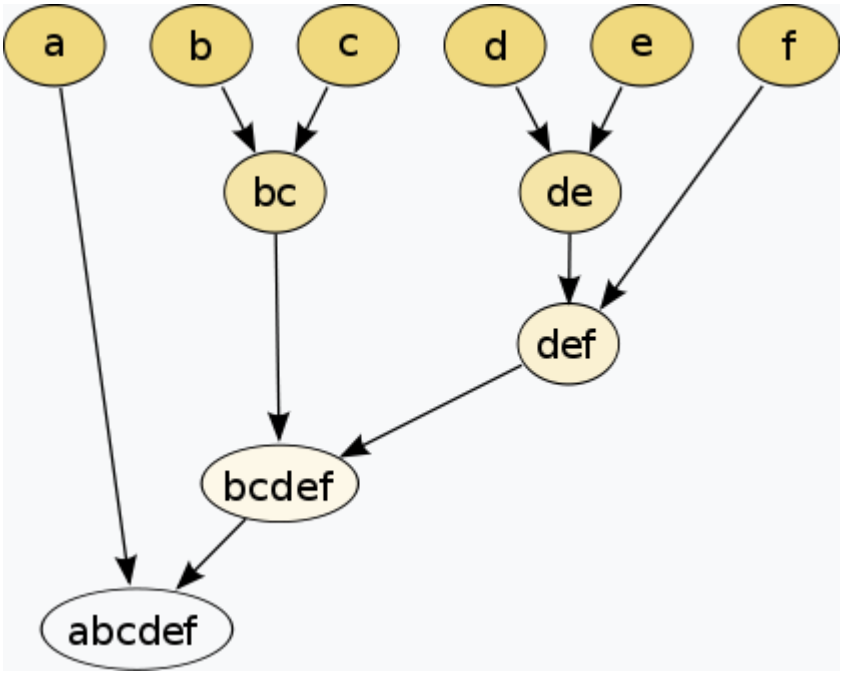
在一个指定高度上切割此树，可以得到一个相应精度的分类。

## 聚集型层次聚类



Raw data

它的层次聚类树如下图



Traditional representation

### 概念聚类

# 分散性聚类

## K-均值法及衍生算法

#### K-均值法聚类

K-均值算法表示以空间中k个点为中心进行聚类，对最靠近他们的对象归类。

> *例如：*数据集合为三维，聚类以两点：$X = (x_1, x_2, x_3), Y = (y_1, y_2, y_3)$。中心点$Z$变为$Z = (z_1, z_2, z_3)$，其中$z_1 = (x_1 + y_1)/2$，$z_2 = (x_2 + y_2)/2$，$z_3 = (x_3 + y_3)/2$。

算法归纳为（J. MacQueen, 1967）：

- 选择聚类的个数*k*.
- 任意产生*k*个聚类，然后确定聚类中心，或者直接生成*k*个中心。
- 对每个点确定其聚类中心点。
- 再计算其聚类新中心。
- 重复以上步骤直到满足收敛要求。（通常就是确定的中心点不再改变。）

该算法的最大优势在于简洁和快速。劣势在于对于一些结果并不能够满足需要，因为结果往往需要随机点的选择非常巧合。

#### QT聚类算法

## 图论方法

# 谱聚类

# 应用

## 生物

## 市场研究

## 其他应用


- Abdi, H. (1994). Additive-tree representations (with an application to face processing). Lecture Notes in Biomathematics, 84, 43-59.. 1990.
- Clatworthy, J., Buick, D., Hankins, M., Weinman, J., & Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British Journal of Health Psychology* 10: 329-358.
- Cole, A. J. & Wishart, D. (1970). An improved algorithm for the Jardine-Sibson method of generating overlapping clusters. *The Computer Journal* 13(2):156-163.
- Ester, M., Kriegel, H.P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA: AAAI Press, pp. 226–231.
- Heyer, L.J., Kruglyak, S. and Yooseph, S., Exploring Expression Data: Identification and Analysis of Coexpressed Genes, *Genome Research* 9:1106-1115.
- Huang, Z. (1998). Extensions to the K-means Algorithm for Clustering Large Datasets with ategorical Values. *Data Mining and Knowledge Discovery*, 2, p. 283-304.
- Jardine, N. & Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *The Computer Journal* 11:177.
- The on-line textbook: Information Theory, Inference, and Learning Algorithms, by David J.C. MacKay includes chapters on k-means clustering, soft k-means clustering, and derivations including the E-M algorithm and the variational view of the E-M algorithm.
- Ng, R.T. and Han, J. 1994. Efficient and effective clustering methods for spatial data mining. Proceedings of the 20th VLDB Conference, Santiago, Chile, pp. 144–155.
- Prinzie A., D. Van den Poel (2006), Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM. *Decision Support Systems* 42 (2): 508-526.
- Romesburg, H. Clarles, *Cluster Analysis for Researchers*, 2004, 340 pp. ISBN 1-4116-0617-5 or publisher, reprint of 1990 edition published by Krieger Pub. Co.. A Japanese language translation is available from Uchida Rokakuho Publishing Co., Ltd., Tokyo, Japan.
- Zhang, T., Ramakrishnan, R., and Livny M. 1996. BIRCH: An efficient data clustering method for very large databases. Proceedings of ACM SIGMOD Conference, Montreal, Canada, pp. 103–114.

For spectral clustering :

- Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888-905, August 2000. Available on Jitendra Malik's homepage
- Marina Meila and Jianbo Shi, "Learning Segmentation with Random Walk", Neural Information Processing Systems, NIPS, 2001. Available from Jianbo Shi's homepage

For estimating number of clusters:

- Can, F., Ozkarahan, E. A. (1990) "Concepts and effectiveness of the cover coefficient-based clustering methodology for text databases." ACM Transactions on Database Systems. 15 (4) 483-517.

For discussion of the elbow criterion:

- Aldenderfer, M.S., Blashfield, R.K., *Cluster Analysis*, (1984), Newbury Park (CA): Sage.


# 外部链接

- P. Berkhin, *Survey of Clustering Data Mining Techniques*, Accrue Software, 2002.
- Jain, Murty and Flynn: *Data Clustering: A Review*, ACM Comp. Surv., 1999.

- for another presentation of hierarchical, k-means and fuzzy c-means see this introduction to clustering. Also has an explanation on mixture of Gaussians.
- David Dowe, *Mixture Modelling page* - other clustering and mixture model links.
- a tutorial on clustering[1]
- The on-line textbook: Information Theory, Inference, and Learning Algorithms, by David J.C. MacKay includes chapters on k-means clustering, soft k-means clustering, and derivations including the E-M algorithm and the variational view of the E-M algorithm.

## 相关软件

### 免费类

- The flexclust package for R
- COMPACT - Comparative Package for Clustering Assessment（in Matlab）
- YALE (Yet Another Learning Environment): freely available open-source software for data pre-processing, knowledge discovery, data mining, machine learning, visualization, etc. also including a plugin for clustering, fully integrating Weka, easily extendible, and featuring a graphical user interface as well as a XML-based scripting language for data mining;
- mixmod : Model Based Cluster And Discriminant Analysis. Code in C++, interface with Matlab and Scilab
- LingPipe Clustering Tutorial Tutorial for doing complete- and single-link clustering using LingPipe, a Java text data mining package distributed with source.
- Weka : Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.
- Tanagra : a free data mining software including several clustering algorithms such as K-MEANS, SOM, Clustering Tree, HAC and more.
- Cluster : Open source clustering software. The routines are available in the form of a C clustering library, an extension module to Python, a module to Perl.
- python-cluster Pure python implementation

### 商业类

- Clustan
- Peltarion Synapse（using self-organizing maps）[2]

---