聚类分析算法

聚类分析算法（Principal component analysis，PCA)）也是一种无监督学习算法，针对没有响应变量而仅有特征变量的数据集，其主要作用就是快速分类。虽然是无监督学习算法，聚类分析也有很多应用场景，比如电商平台系统针对具有相似购买行为的用户进行聚类，针对划分好的客户类别，将某用户购买的产品在同一类别用户内相互推荐，实现精准促销；或者根据以往销售记录及其他特征对产品进行聚类，若某用户购买了一款产品，则继续向其推送同一类别内其他产品。又比如大型连锁企业依据销售情况对全辖门店进行有效分类，根据聚类结果制定生产计划、开展物流配送、实现更有效率的资源配置等。

目录

CONTENTS

1 聚类分析算法的基本原理
2 数据准备
3 划分聚类分析算法示例
4 层次聚类分析算法示例
5 习 题

PART 01

聚类分析算法的基本原理

聚类分析算法的基本原理

聚类分析是根据特征变量，按照一定的标准对样本示例进行分类。

按照分析方法的不同，聚类分析分成两个宽泛的类别，包括划分聚类分析和层次聚类分析。划分聚类分析是一种快速聚类方式，它将数据看作K维空间上的点，以距离为标准进行聚类分析，将样本分为指定的K类，包括K个平均数的聚类分析方法、K个中位数的聚类分析方法。划分聚类分析过程只限于连续数据，要求预先指定聚类数目。

层次聚类分析也称系统聚类分析，基本思路是对相近程度最高的两类进行合并，组成一个新类并不断重复此过程，直到所有的个体都归为一类，通常只限于较小的数据文件（要聚类的对象只有数百个）。

划分聚类分析

划分聚类分析方法的基本思想是将观测到的样本划分到一系列事先设定好的不重合的分组中去。划分聚类分析方法主要包括两种：一种是K个平均数的聚类分析方法，此方法的操作流程是通过迭代过程将样本示例分配到具有最接近的平均数的组，然后找出这些聚类；另一种是K个中位数的聚类分析方法，此方法的操作流程是通过迭代过程将样本示例分配到具有最接近的中位数的组，然后找出这些聚类。

以K均值聚类分析为例，K均值聚类分析的基本原理是：首先指定聚类的个数并按照一定的规则选取初始聚类中心，让个案向最近的聚类中心靠拢，形成初始分类，然后按最近距离原则不断修改不合理分类，直至合理为止。

划分聚类分析

比如用户选择x个特征变量参与聚类分析，最后要求聚类数为y，那么将由系统首先选择y个样本示例（当然也可由用户指定）作为初始聚类中心，x个特征变量组成x维特征空间。每个样本示例在x维特征空间中是一个点，y个事先选定的样本示例案就是y个初始聚类中心点。然后系统按照距这几个初始聚类中心距离最小的原则把样本示例分派到各类中心所在的类中去，构成第一次迭代形成的y类。

然后系统根据组成每一类的样本示例，计算各特征变量均值，每一类中的x个特征均值在x维特征空间中又形成y个点，这就是第二次迭代的聚类中心，按照这种方法依次迭代下去，直到达到指定的迭代次数或达到终止迭代的要求时，迭代停止，形成最终聚类中心。

划分聚类分析

K-均值聚类法计算量小、占用内存少并且处理速度快，因此比较适合处理大样本的聚类分析。

划分聚类分析方法与层次聚类分析方法相比在计算上相对简单且计算速度更快一些，但是它也有自己的缺点，它要求事先指定样本聚类的精确数目，这与聚类分析探索性的本质是不相适应的。

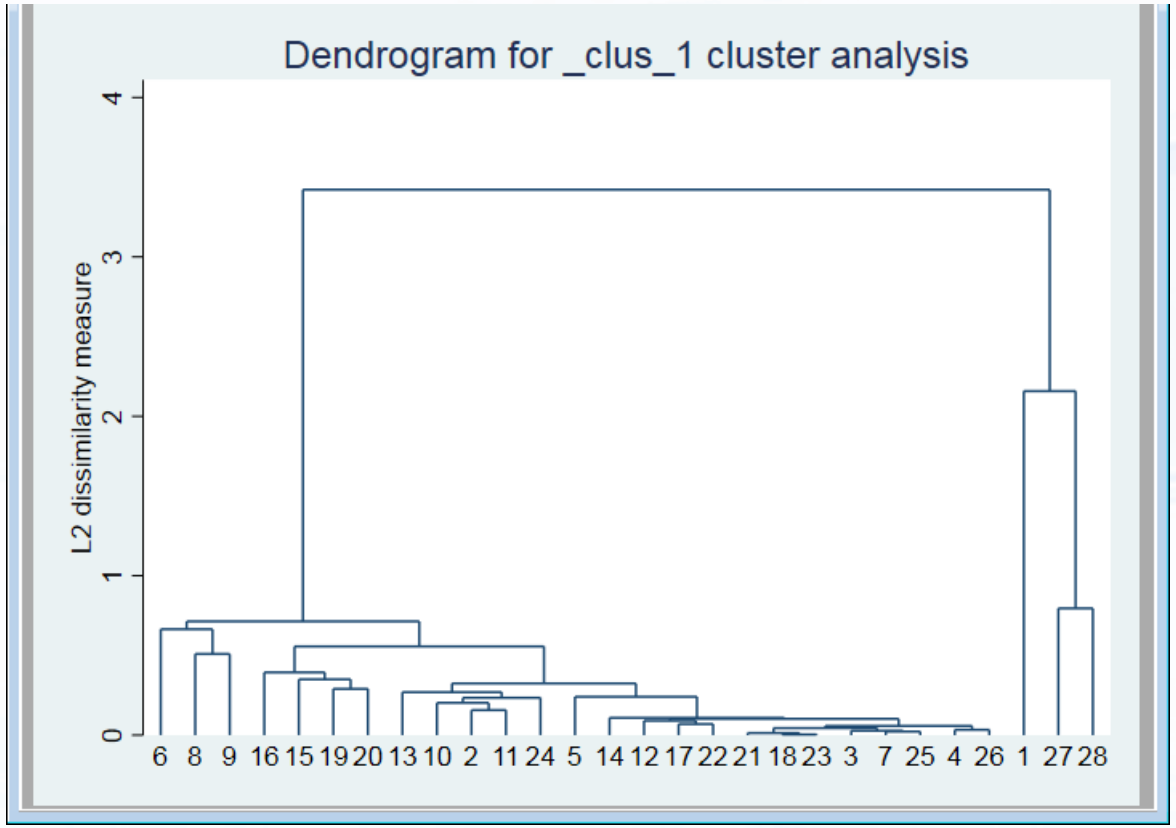
层次聚类分析

层次聚类分析也称系统聚类分析，与划分聚类分析方法的原理不同，层次聚类分析的基本原理是根据选定的特征来识别相对均一的个案（变量）组，使用的算法是开始将每个个案（或变量）都视为一类，然后根据类与类之间的距离或相似程度将最近的类加以合并，再计算新类与其他类之间的相似程度，并选择最相似的加以合并，这样每合并一次就减少一类，不断继续这一过程，最终实现完全聚类，即把所有的观测样本汇集到一个组中。

层次聚类分析

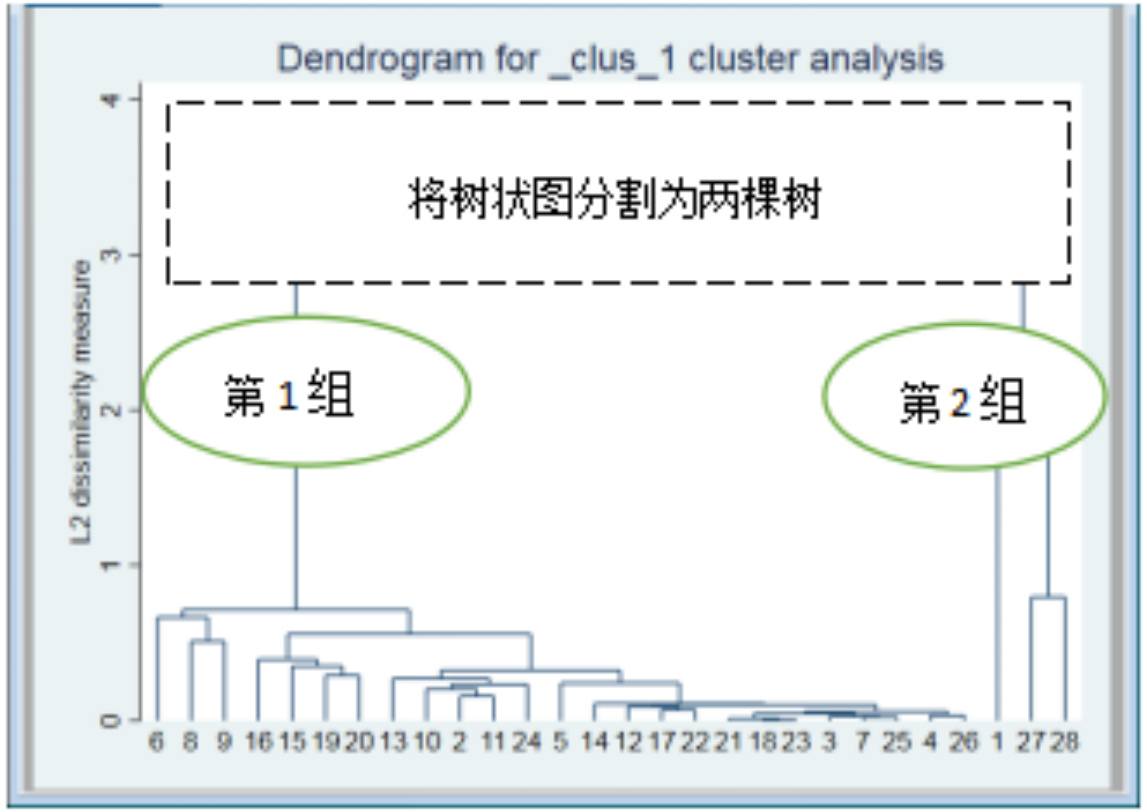
在实际分析中常用到的一个工具是树状图。图12.1即为树状图的一个示例。（图片来源：《Stata统计分析从入门到精通》杨维忠、张甜著，清华大学出版社，第12章聚类分析）。

观察该图，可以直观地看到具体的聚类情况，如21、18、23号样本聚合在一起（当然图片放大后，可能会进一步细分，比如21、18、23号样本首先是18、23号合并，然后与21合并），3、7、25号样本聚合在一起。



层次聚类分析

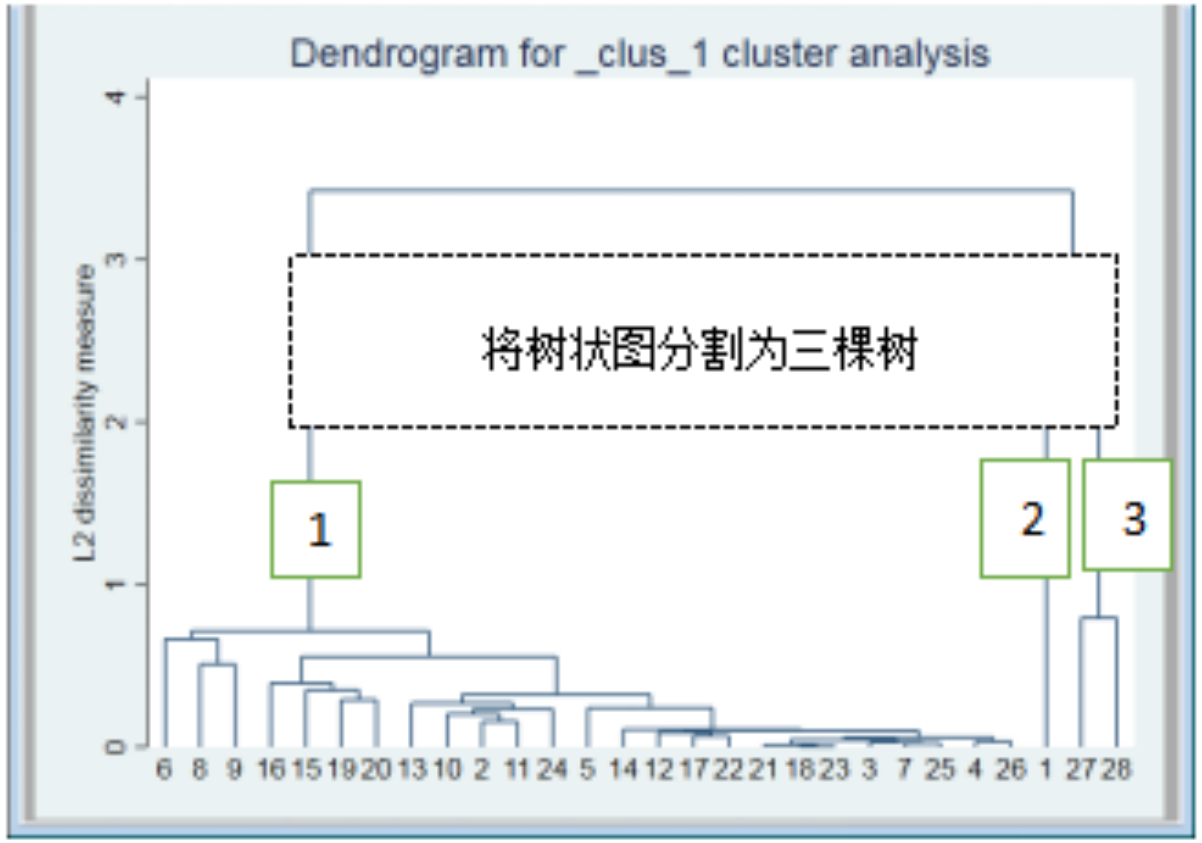
那么，到底分成了多少类呢？取决于研究的需要和实际的情况，需要用户加入自己的判断，确定好分类需求后，从聚类分析树状图最上面使用分割框往下进行分割。例如需要把所有样本观测值分成2类时，分析结果如图12.2所示。



层次聚类分析

可以发现1、27、28三个样本为第2类，其他样本为第1类。如果需要把所有样本观测值分成3类，就需要自上而下继续进行分割（图中的分割框下移），分析结果如图12.3所示。

可以发现1号样本为第2类，27号、28号样本为第3类，其他样本为第1类。



层次聚类分析

与划分聚类分析方法相比，层次聚类分析方法的计算过程更为复杂，计算速度相对较慢，但是它不要求事先指定需要分类的数量，这一点是符合聚类分析探索性的本质特点的，所以这种聚类分析方法应用也非常广泛。

氐□□□技鳖□颃□	□蔽□箔□□□□□掊□颃咙
跬□□□□□□技鳖 ˚Single-Linkage Cluster Analysis `	帜搏□蔽□箔□□□□柄变□表跬□□□貔□涤箔□□□□Y
跬□□□□□□技鳖 ˚Complete-Linkage Cluster Analysis `	帜搏□蔽□箔□□□□柄变□表跬□□□貔□涤箔□□□□Y
薏寿□□□□□技鳖 ˚Average-Linkage Cluster Analysis `	帜搏□蔽□箔□□□□柄变□表龄□貔□涤□薏寿竦箔□□□□Y
表大竦□□□□□技鳖 ˚Median-Linkage Cluster Analysis `	帜搏□蔽□箔□□□□柄变□表龄□貔□涤□表大竦箔□□□□Y
□微□□□□□技鳖 ˚Centroid-Linkage Cluster Analysis `	帜搏□蔽□箔□□□□柄变□表龄□貔□涤□□微箔□□□□Y
ward □□□□□技鳖 ˚ward-Linkage Cluster Analysis `	变标□□箔□□□□□拆□兼□徼□粉莞薏颃冒 ！登□瑚龄□貔□涤□粉莞薏颃冒跬综 ！□□粉莞薏颃冒匕料□香 Y

衡量样本示例距离的测度

无论是划分聚类分析还是层次聚类分析，聚类分析的本质都是按照一定的距离对样本示例进行分割，在一定标准距离范围以内的样本示例被划分为一个类别，所以如何测度样本数据之间的距离也非常重要。

常用的针对连续变量数据的标准如表所示。

简写	含　义	适用变量类型
Euclidean ·L2·	□□□□Y□□恨啡柄龄□啡箔□薏颃莞箔冒□薏颃□ $\mathrm{dist}(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$	□□粮□竦腌
L2squared	□□□□□薏颃	□□粮□竦腌
Manhattan ·L1·	跎渺□□□ !擦□□着啡□□	□□粮□竦腌
Linfinity	跬香啡□□	□□粮□竦腌
correlation	□哄□竦□粹忸掊□	□□粮□竦腌
L(#)	蝓料乡耶绦□□ !□□屎咎磷瑚贽帜 p 啡 Y □□恨啡柄龄□啡箔□ p □葸□着莞箔冒□ p □□ $d_{xy}=\left(\sum_{u=1}^{n}\lvert x_u-y_u\rvert^{p}\right)^{\frac{1}{p}}$ p 啡炼啡□射柄 1~7Y 嘀 p=1 蜉 !丐蟾』书〖·跎渺□□□ 丬嘀 p=2 蜉 !丐蟾□啡□□丬嘀 p □□拆蚯□香蜉 !丐蟾冀□□乡□□	□□粮□竦腌
cosine	怠喋□粹掊	□□粮□竦腌

衡量样本示例距离的测度

常用的针对分类变量数据的标准如表所示。

针对聚类分析算法，同样要求首先对特征变量数据进行标准化，之所以这么做，是因为如果不进行标准化，少数变量可能会对距离的影响太大，使得分析结果严重失真。

简写	含　义	适用变量类型
matching	□狂空□□竦	技□粮□竦腌
Jaccard	Jaccard □粹□竦	技□粮□竦腌
Russell	Russell 冒 Rao □粹□竦	技□粮□竦腌
Hamann	Hamann □粹□竦	技□粮□竦腌
Dice	Dice □粹□竦	技□粮□竦腌
antiDice	镰 Dice □粹□竦	技□粮□竦腌
Sneath	Sneath 冒 Sokal □粹□竦	技□粮□竦腌
Rogers	Rogers 冒 Tanimoto □粹□竦	技□粮□竦腌
Ochiai	Ochiai □粹□竦	技□粮□竦腌
Yule	Yule □粹□竦	技□粮□竦腌
Anderberg	Anderberg □粹□竦	技□粮□竦腌
Kulczynski	Kulczynski □粹□竦	技□粮□竦腌
Pearson	Pearson □粹□竦	技□粮□竦腌
Gower2	蔽 Pearson □领技□□□粹□竦	技□粮□竦腌

PART 02

数据准备

数据准备

本节我们以“数据12.1”为例进行讲解，其中的数据为《中国2019年分地区连锁餐饮企业基本情况统计》，摘编自《中国统计年鉴2020》。这个数据文件中共有9个变量，分别是V1~V9，分别表示地区、总店数、门店总数、年末从业人数、年末餐饮营业面积、餐位数、营业额、商品购进总额、统一配送商品购进额。

下面我们以V3门店总数、V6餐位数、V7营业额、V8商品购进总额4个变量对所有样本观测值开展划分聚类分析。

载入分析所需要的模块和函数

在进行分析之前，我们首先载入分析所需要的模块和函数，读取数据集并进行观察。

示例

参阅教材内容

变量设置及数据处理

示例

参阅教材内容

特征变量相关性分析

示例

参阅教材内容

PART 03

划分聚类分析算法示例

使用K均值聚类分析方法对样本示例进行聚类(K=2)

示例

参阅教材内容

使用K均值聚类分析方法对样本示例进行聚类(K=3)

示例

参阅教材内容

使用K均值聚类分析方法对样本示例进行聚类(K=4)

示例

参阅教材内容

PART 04

层次聚类分析算法示例

最短联结法聚类分析

示例

参阅教材内容

最长联结法聚类分析

示例

参阅教材内容

平均联结法聚类分析

示例

参阅教材内容

ward联结法聚类分析

示例

参阅教材内容

重心联结法聚类分析

示例

参阅教材内容

PART 05

习　题

习题

继续使用“数据12.1”，以总店数、年末从业人数、年末餐饮营业面积、统一配送商品购进额4个变量，即V2、V4、V5、V9，对所有样本观测值开展划分聚类分析和层次聚类分析。

（1）载入分析所需要的库和模块

（2）变量设置及数据处理

（3）特征变量相关性分析

（4）使用K均值聚类分析方法对样本示例进行聚类(K=2)

（5）使用K均值聚类分析方法对样本示例进行聚类(K=3)

习题

（6）使用K均值聚类分析方法对样本示例进行聚类(K=4)

感谢聆听

THANKS