聚类理论及其 R 软件实现

黄放（人大经济论坛· 数据处理部）

聚类分析（Cluster Analysis）又称群分析，是根据“物以类聚”的道理，对样品或指标进行分类的一种多元统计分析方法，它们讨论的对象是大量的样品，要求能合理地按各自的特性来进行合理的分类，没有任何模式可供参考或依循，即是在没有先验知识的情况下进行的。聚类分析起源于分类学，在古老的分类学中，人们主要依靠经验和专业知识来实现分类，很少利用数学工具进行定量的分类。随着人类科学技术的发展，对分类的要求越来越高，以致有时仅凭经验和专业知识难以确切地进行分类，于是人们逐渐地把数学工具引用到了分类学中，形成了数值分类学，之后又将多元分析的技术引入到数值分类学形成了聚类分析。

聚类分析被应用于很多方面，在商业上，聚类分析被用来发现不同的客户群，并且通过购买模式刻画不同的客户群的特征；在生物上，聚类分析被用来动植物分类和对基因进行分类，获取对种群固有结构的认识；在地理上，聚类能够帮助在地球中被观察的数据库商趋于的相似性；在保险行业上，聚类分析通过一个高的平均消费来鉴定汽车保险单持有者的分组，同时根据住宅类型，价值，地理位置来鉴定一个城市的房产分组；在因特网应用上，聚类分析被用来在网上进行文档归类来修复信息。

聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。聚类分析的目标就是在相似的基础上收集数据来分类。聚类源于很多领域，包括数学，计算机科学，统计学，生物学和经济学。在不同的应用领域，很多聚类技术都得到了发展，这些

技术方法被用作描述数据，衡量不同数据源间的相似性，以及把数据源分类到不同的簇中。

主要的几种聚类算法

聚类分析计算方法主要有如下几种：分裂法(partitioning methods)：层次法(hierarchical methods)：基于密度的方法(density-based methods)：基于网格的方法(grid-basedmethods)：基于模型的方法(model-based methods)。

1. 分裂法又称划分方法(PAM:PArtitioning method) 首先创建 k 个划分，k 为要创建的划分个数；然后利用一个循环定位技术通过将对象从一个划分移到另一个划分来帮助改善划分质量。
   - k-means,k-medoids,CLARA(Clustering LARge Application),
   - CLARANS(Clustering Large Application based upon RANdomized Search).
2. 层次法(hierarchical method) 创建一个层次以分解给定的数据集。该方法可以分为自上而下（分解）和自下而上（合并）两种操作方式。为弥补分解与合并的不足，层次合并经常要与其它聚类方法相结合，如循环定位。
   - BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies) 方法，它首先利用树的结构对对象集进行划分；然后再利用其它聚类方法对这些聚类进行优化。
   - CURE(Clustering Using REprisentatives) 方法，它利用固定数目代表对象来表示相应聚类；然后对各聚类按照指定量（向聚类中心）进行收缩。
   - ROCK 方法，它利用聚类间的连接进行聚类合并。
   - CHEMALOEN 方法，它则是在层次聚类时构造动态模型。
3. 基于密度的方法，根据密度完成对象的聚类。它根据对象周围的密度（如 DBSCAN）不断增长聚类。
   - DBSCAN(Densit-based Spatial Clustering of Application with Noise):该算法通过不断生长足够高密度区域来进行聚类；它能从含有噪声的空间数据库中发现任意形状的聚类。此方法将一个聚类定义为一组“密度连接”的点集。
   - OPTICS(Ordering Points To Identify the Clustering Structure):并不明确产

生一个聚类，而是为自动交互的聚类分析计算出一个增强聚类顺序。

4. 基于网格的方法，首先将对象空间划分为有限个单元以构成网格结构；然后利用网格结构完成聚类。
   - STING(STatistical INformation Grid) 就是一个利用网格单元保存的统计信息进行基于网格聚类的方法。
   - CLIQUE(Clustering In QUEst)和 Wave-Cluster 则是一个将基于网格与基于密度相结合的方法。
5. 基于模型的方法，它假设每个聚类的模型并发现适合相应模型的数据。
   - 统计方法 COBWEB:是一个常用的且简单的增量式概念聚类方法。它的输入对象是采用符号量（属性-值）对来加以描述的。采用分类树的形式来创建一个层次聚类。
   - CLASSIT 是 COBWEB 的另一个版本.。它可以对连续取值属性进行增量式聚类。它为每个结点中的每个属性保存相应的连续正态分布（均值与方差）；并利用一个改进的分类能力描述方法，即不象 COBWEB 那样计算离散属性（取值）和而是对连续属性求积分。但是 CLASSIT 方法也存在与 COBWEB 类似的问题。因此它们都不适合对大数据库进行聚类处理.

传统的聚类算法已经比较成功的解决了低维数据的聚类问题。但是由于实际应用中数据的复杂性，在处理许多问题时，现有的算法经常失效，特别是对于高维数据和大型数据的情况。因为传统聚类方法在高维数据集中进行聚类时，主要遇到两个问题。①高维数据集中存在大量无关的属性使得在所有维中存在簇的可能性几乎为零；②高维空间中数据较低维空间中数据分布要稀疏，其中数据间距离几乎相等是普遍现象，而传统聚类方法是基于距离进行聚类的，因此在高维空间中无法基于距离来构建簇。

高维聚类分析已成为聚类分析的一个重要研究方向。同时高维数据聚类也是聚类技术的难点。随着技术的进步使得数据收集变得越来越容易，导致数据库规模越来越大、复杂性越来越高，如各种类型的贸易交易数据、Web 文档、基因表达数据等，它们的维度（属性）通常可以达到成百上千维，甚至更高。但是，受“维度效应”的影响，许多在低维数据空间表现良好的聚类方法运用在高维空

间上往往无法获得好的聚类效果。高维数据聚类分析是聚类分析中一个非常活跃的领域，同时它也是一个具有挑战性的工作。目前，高维数据聚类分析在市场分析、信息安全、金融、娱乐、反恐等方面都有很广泛的应用。另外，聚类分析是根据事物本身的特性研究个体的一种方法，目的在于将相似的事物归类。它的原则是同一类中的个体有较大的相似性，不同类的个体差异性很大。这种方法有三个特征：（1）适用于没有先验知识的分类。如果没有这些事先的经验或一些国际标准、国内标准、行业标准，分类便会显得随意和主观。这时只要设定比较完善的分类变量，就可以通过聚类分析法得到较为科学合理的类别； 可以处理多个变量决定的分类。例如，要根据消费者购买量的大小进行分类比较容易，但如果在进行数据挖掘时，要求根据消费者的购买量、家庭收入、家庭支出、年龄等多个指标进行分类通常比较复杂，而聚类分析法可以解决这类问题； （3）聚类分析法是一种探索性分析方法，能够分析事物的内在特点和规律，并根据相似性原则对事物进行分组，是数据挖掘中常用的一种技术。

系统聚类法以及 R 语言的实现程序

系统聚类法（hierarchical clustering method）是聚类分析诸方法中用得最多的一种，其中基本的思想是：开始将 N 个样本各自作为一类，并规定样本之间的距离和类与类之间的距离和类与类之间的距离，然后将距离最近的两类合并成一个新类，并计算新类与其他类的距离；重复进行两个最近类的合并，每次减少一类，直至所有的样本合成一类。

以下用 d_{ij} 表示第 i 个样本与第 j 个样本的距离，G_1，G_2……表示类，D_{KL} 表示 G_K 与 G_L 的距离。在下面的所介绍的系统聚类法中，所有的方法一开始每个样本自成一类，类鱼类之间的距离与样本之间的距离相同，即 $D_{KL}=d_{KL}$，所以最初的距离矩阵全部相同，记为 $D_{(0)}=(d_{ij})$。

- 最短距离法

定义类与类之间的距离为两类最近样本间的距离，即

$$D_{KL}=\min_{i\in G_K,\, j\in G_L} d_{ij}$$

称这种系统聚类法最短聚类法（single linkage method）.

当某步骤类 G_K 和 G_L 合并为 G_M 后，按最短距离法计算新类与其它类 G_J 的类

间公式，其递推公式为：

$$D_{KL}=\min_{i\in G_K,j\in G_L} d_{ij}=\min\{\min_{i\in G_K,j\in G_L} d_{ij},\min_{i\in G_L,j\in G_J} d_{ij}\}=\min\{D_{KL},D_{LJ}\}$$

其 R 语言的程序是：

Hcluster(d,"single")

- 最大距离法

定义类与类之间的距离为两类最近样本间的距离，即

$$D_{KL}=\max_{i\in G_K,j\in G_L} d_{ij}$$

称这种系统聚类法最长聚类法（complete linkage method）。

当某步骤类 G_K 和 G_L 合并为 G_M 后，则 G_M 与任一类 G_J 距离为

$$D_{KL}=\max\{D_{KL},D_{LJ}\}$$

其 R 语言的程序是：

Hcluster(d,"complete")

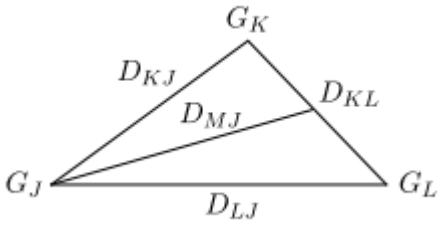
- 中间距离法

类与类之间的距离几不取两类最近样本的距离，也不去两类最远样本的距离，而是取介于两者中间的距离，称为中间距离法（median method）

设某一步将 G_K 和 G_L 合并为 G_M，对于任意类 G_J，考虑 D_{KL}，D_{LJ} 和 D_{KJ} 由和为边长组成的三角形，取变得中线 D_{KL} 作为 G_{MJ}，由初等平面几何可知，的计算公式为

$$D_{MJ}^2=\frac{1}{2}D_{KJ}^2+\frac{1}{2}D_{LJ}^2-\frac{1}{4}D_{KL}^2$$

这就是中间距离法的递推公式。



中间法可推广位更一般的情形，将下式中三个系数改为带有参数的 β，即

$$D_{MJ}^2=\frac{1-\beta}{2}(D_{KJ}^2+D_{LJ}^2)-\beta D_{KL}^2$$

其中 β <1，这种方法成为可变法。当 β =0 时，递推公式变为

$$D_{MJ}^2 = \frac{1}{2}(D_{KJ}^2 + D_{LJ}^2)$$

称此方法为 Mcquitty 相似分析法。

其 R 语言的程序是：

Hcluster(d,"mcquitty")

- 离差平方和

离差平方和时 Ward(1936)提出的，也称为 Ward 法。它基于方差分析的思想，如果类分得正确，则同类样本之间的离差平和应当较小，不同样本之间的离差平均和应当较大。设类 G_K 和 G_L 合并为新类 G_M，则 G_K，G_L 和 G_M 的离差平方和分别是

$$W_K = \sum_{i \in G_K} (x_{(i)} - \bar{x}_K)^T (x_{(i)} - \bar{x}_K)$$

$$W_L = \sum_{i \in G_L} (x_{(i)} - \bar{x}_L)^T (x_{(i)} - \bar{x}_L)$$

$$W_M = \sum_{i \in G_M} (x_{(i)} - \bar{x}_M)^T (x_{(i)} - \bar{x}_M)$$

其中 $\bar{x}_K$， $\bar{x}_L$ 和 $\bar{x}_M$ 分别是 G_K， G_L 和 G_M 的重心。所以 W_K， W_L 和 W_M 反映了各自类内样本的分散程度。如 G_K 和 G_L 这两类相聚较近，则合并后所增加的离差平方和 W_M -W_K -W_L 应较小；否这应较大。于是定义 G_K 和 G_L 之间的平方距离为

$$D^2{}_{KL} = \frac{n_K n_L}{n_M}(\bar{x}K - \bar{x}L)^T(\bar{x}K - \bar{x}L)$$

这种系统聚类法称之为离差平和法或 Ward 方法（Ward's mininum variance method）。它的递推公式为

$$D^2{}_{ML} = \frac{n_J + n_K}{n_J + n_M} D^2{}_{KL} + \frac{n_J + n_L}{n_J + n_M} D^2{}_{LJ} - \frac{n_J}{n_J + n_M} D^2{}_{KL}$$

G_K 和 G_L 之间的平方距离也可以写成

$$D^2_{KL} = \frac{n_K n_L}{n_M}(\bar{x}K - \bar{x}L)^T(\bar{x}K - \bar{x}L)$$

可见，这个距离与重心法的距离只相差了一个常数倍。重心法的类间距与两类的样本数无关，而离差平方和法的类间距与两类的样本数由较大的关系，两个大类倾向于由较大的距离，因而不易合并，这更符合对聚类的实际要求。离差平方和在许多场合下由于重心法，是一种比较好的系统聚类法，但它对异常值很敏感。

其 R 语言的程序是：

Hcluster(d,"ward")

聚类方法的应用

这种较成熟的统计学方法如果在市场分析中得到恰当的应用，必将改善市场营销的效果，为企业决策提供有益的参考。其应用的步骤为：将市场分析中的问题转化为聚类分析可以解决的问题，利用相关软件（如 SPSS、SAS 等）求得结果，由专家解读结果，并转换为实际操作措施，从而提高企业利润，降低企业成本。

聚类分析在客户细分中的应用

消费同一种类的商品或服务时，不同的客户有不同的消费特点，通过研究这些特点，企业可以制定出不同的营销组合，从而获取最大的消费者剩余，这就是客户细分的主要目的。常用的客户分类方法主要有三类：经验描述法，由决策者根据经验对客户进行类别划分；传统统计法，根据客户属性特征的简单统计来划分客户类别；非传统统计方法，即基于人工智能技术的非数值方法。聚类分析法兼有后两类方法的特点，能够有效完成客户细分的过程。

例如，客户的购买动机一般由需要、认知、学习等内因和文化、社会、家庭、小群体、参考群体等外因共同决定。要按购买动机的不同来划分客户时，可以把前述因素作为分析变量，并将所有目标客户每一个分析变量的指标值量化出来，再运用聚类分析法进行分类。在指标值量化时如果遇到一些定性的指标值，可以用一些定性数据定量化的方法加以转化，如模糊评价法等。除此之外，可以将客户满意度水平和重复购买机会大小作为属性进行分类；还可以在区分客户之间差异性的问题上纳入一套新的分类法，将客户的差异性变量划分为五类：产品利益、

客户之间的相互作用力、选择障碍、议价能力和收益率，依据这些分析变量聚类得到的归类，可以为企业制定营销决策提供有益参考。

以上分析的共同点在于都是依据多个变量进行分类，这正好符合聚类分析法解决问题的特点；不同点在于从不同的角度寻求分析变量，为某一方面的决策提供参考，这正是聚类分析法在客户细分问题中运用范围广的体现。

聚类分析在实验市场选择中的应用

实验调查法是市场调查中一种有效的一手资料收集方法，主要用于市场销售实验，即所谓的市场测试。通过小规模的实验性改变，以观察客户对产品或服务的反应，从而分析该改变是否值得在大范围内推广。

实验调查法最常用的领域有：市场饱和度测试。市场饱和度反映市场的潜在购买力，是市场营销战略和策略决策的重要参考指标。企业通常通过将消费者购买产品或服务的各种决定因素（如价格等）降到最低限度的方法来测试市场饱和度。或者在出现滞销时，企业投放类似的新产品或服务到特定的市场，以测试市场是否真正达到饱和，是否具有潜在的购买力。前述两种措施由于利益和风险的原因，不可能在企业覆盖的所有市场中实施，只能选择合适的实验市场和对照市场加以测试，得到近似的市场饱和度；产品的价格实验。这种实验往往将新定价的产品投放市场，对顾客的态度和反应进行测试，了解顾客对这种价格的是否接受或接受程度；新产品上市实验。波士顿矩阵研究的企业产品生命周期图表明，企业为了生存和发展往往要不断开发新产品，并使之向明星产品和金牛产品顺利过渡。然而新产品投放市场后的失败率却很高，大致为 66%到 90%。因而为了降低新产品的失败率，在产品大规模上市前，运用实验调查法对新产品的各方面（外观设计、性能、广告和推广营销组合等）进行实验是非常有必要的。

在实验调查方法中，最常用的是前后单组对比实验、对照组对比实验和前后对照组对比实验。这些方法要求科学的选择实验和非实验单位，即随机选择出的实验单位和非实验单位之间必须具备一定的可比性，两类单位的主客观条件应基本相同。

通过聚类分析，可将待选的实验市场（商场、居民区、城市等）分成同质的几类小组，在同一组内选择实验单位和非实验单位，这样便保证了这两个单位之间具有了一定的可比性。聚类时，商店的规模、类型、设备状况、所处的地段、

管理水平等就是聚类的分析变量。聚类分析在抽样方案设计中的应用

抽样设计是市场调查中非常重要的一个部分，它的合理性直接决定了市场调查结果的可信度。在抽样方案设计的步骤中，抽样组织形式的选择又是一个关键环节，它决定了样本对总体的代表性的高低。依据抽样误差由低到高的顺序排列，按照标志排队的等距抽样方式抽样误差最小，其次分别为分层抽样、按照无关标志排队的等距抽样、简单随机抽样、整群抽样和非随机抽样。结合资源的限制和操作的方便性进行综合选择，分层抽样在实践中的应用最为广泛。分层抽样又称类型抽样，它是先将总体所有单位按照重要标志进行分组，然后在各组内按照简单随机抽样或等距抽样方式抽取样本单位的一种抽样方式。在分组时引入聚类方法，可以增强组别的合理性。

聚类分析在销售片区确定中的应用

销售片区的确定和片区经理的任命在企业的市场营销中发挥着重要的作用。只有合理地将企业所拥有的子市场归成几个大的片区，才能有效地制定符合片区特点的市场营销战略和策略，并任命合适的片区经理。聚类分析在这个过程中的应用可以通过一个例子来说明。某公司在全国有 20 个子市场，每个市场在人口数量、人均可支配收入、地区零售总额、该公司某种商品的销售量等变量上有不同的指标值。以上变量都是决定市场需求量的主要因素。把这些变量作为聚类变量，结合决策者的主观愿望和相关统计软件提供的客观标准，接下来就可以针对不同的片区制定合理的战略和策略，并任命合适的片区经理了。

聚类分析在市场机会研究中的应用

企业制定市场营销战略时，弄清在同一市场中哪些企业是直接竞争者，哪些是间接竞争者是非常关键的一个环节。要解决这个问题，企业首先可以通过市场调查，获取自己和所有主要竞争者在品牌方面的第一提及知名度、提示前知名度和提示后知名度的指标值，将它们作为聚类分析的变量，这样便可以将企业和竞争对手的产品或品牌归类。根据归类的结论，企业可以获得如下信息：企业的产品或品牌和哪些竞争对手形成了直接的竞争关系。通常，聚类以后属于同一类别的产品和品牌就是所分析企业的直接竞争对手。在制定战略时，可以更多的运用“红海战略”。在聚类以后，结合每一产品或品牌的多种不同属性的研究，可以发现哪些属性组合目前还没有融入产品或品牌中，从而寻找企业在市场中的机

会，为企业制定合理的“蓝海战略”提供基础性的资料。

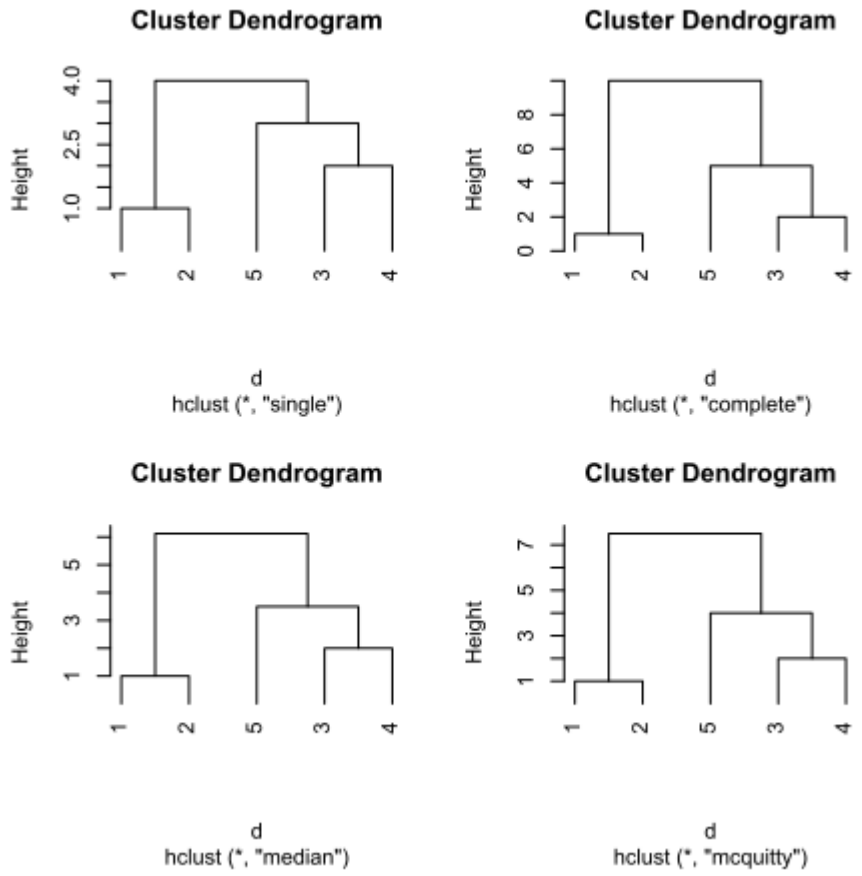
R 语言实现聚类算法的程序

R 语言对于聚类统计分析有着很好的处理效果，例如下图就是正对于四种系统聚类的分析，具体的程序代码以及效果图如下：

以下是 R 语句 (程序名： exam0806.R)

```
#### 输入数据，生成距离结构
x<-c(1,2,6,8,11); dim(x)<-c(5,1); d<-dist(x)
#### 生成系统聚类
hc1<-hclust(d, "single"); hc2<-hclust(d, "complete")
hc3<-hclust(d, "median"); hc4<-hclust(d, "mcquitty")
#### 绘出所有树形结构图，并以 2×2 的形式绘在一张图上
opar <- par(mfrow = c(2, 2))
plot(hc1,hang=-1); plot(hc2,hang=-1)
plot(hc3,hang=-1); plot(hc4,hang=-1)
par(opar)
```

其图形如下：



四种不同距离的谱系图

参考文献：

【1】 邓聚龙 《灰色控制系统》 华中理工大学出版社 1985
【2】 李万绪 “基于灰色关联度的聚类分析方法及其应用”《系统工程》1990，第三期
【3】 人大经济论坛 http://www.pinggu.org
【4】 http://wiki.mbalib.com/wiki/%E8%81%9A%E7%B1%BB%E5%88%86%E6%9E%90
【5】 许云飞 灰色聚类分析方法介绍
【6】 薛毅 统计建模与 R 软件