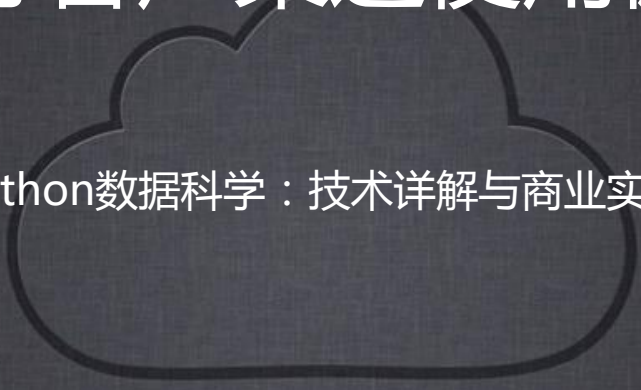


第14章：银行客户渠道使用偏好洞察案例



《Python数据科学：技术详解与商业实践》

讲师：Ben

自我介绍

- 天善商业智能和大数据社区 讲师 –Ben
- 天善社区 ID - Ben_Chang
- <https://www.hellobi.com> – 学习过程中有任何相关的问题都可以提到技术社区数据挖掘版块。

- 客户画像与标签体系
- 客户细分
- 聚类的基本逻辑
- 系统聚类
- K-means聚类
- 使用决策树做聚类后客户分析

1、客户画像与标签体系



什么是客户画像

什么是画像？

客户画像是根据用户社会属性、生活习惯和消费行为等信息而抽象出的一个标签化的用户模型。

构建用户画像的核心工作即是给用户贴“标签”，而标签是通过对用户信息分析而来的高度精炼的特征标识。

为什么要做画像？

精准营销和个性化推荐。

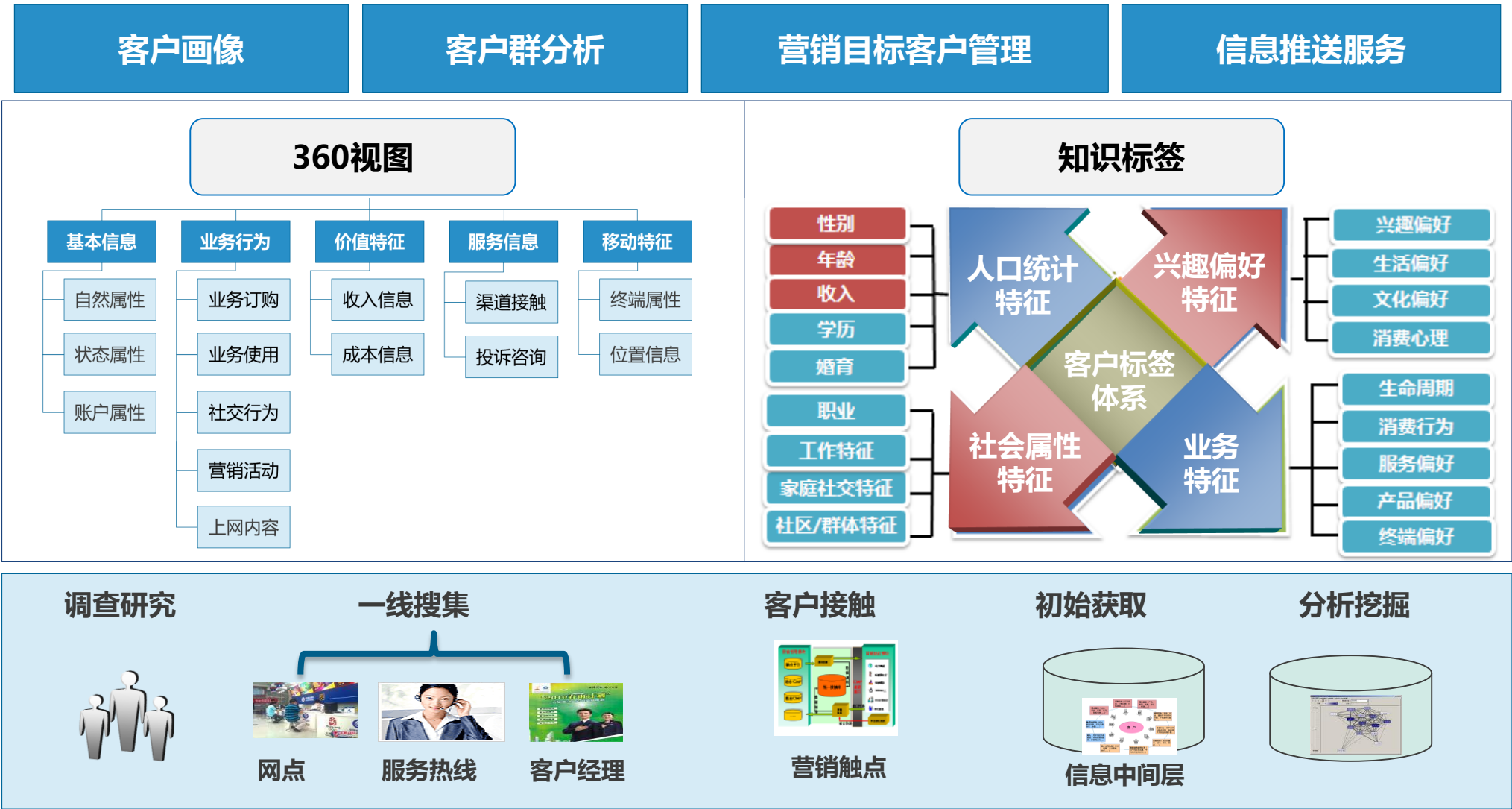
用户统计和理解。

业务经营分析。

怎么做画像？

基本流程：数据收集、行为建模、构建画像。

客户画像与360视图和标签体系



调查研究

一线搜集

客户接触

初始获取

分析挖掘

网点

服务热线

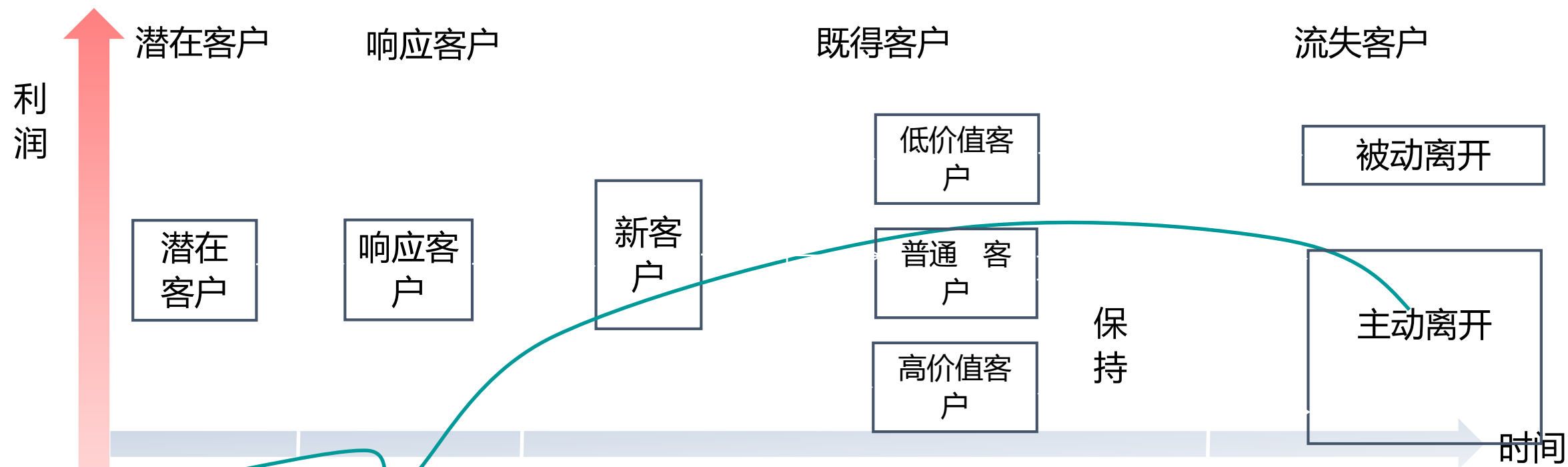
客户经理

营销触点

信息中间层

2、客户细分

客户细分的在客户智能中的位置



●发掘潜在客户

- 客户获取
- 初始信用评分
- 客户价值预测

●客户细分

- 交叉销售
- 产品精准营销
- 行为信用评分
- 欺诈侦测
- 客户保留

什么是客户细分

什么是细分？

将现有消费者群体按一定规则分成若干个小群组, 使得:

- 每一群组的特征描述丰富详细, 且每组都各有不同
 - 这些分组的特征可以引导营销策略制定
- 分群策略使用不同维度的数据
分群策略可用于公司层面制定品牌战略目标,
方便跨部门沟通 (市场, 销售, 服务)

为什么要细分？

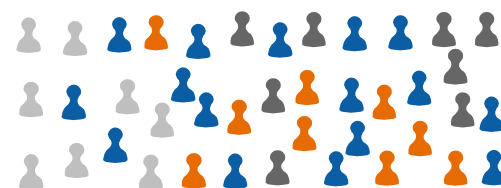
提升销售效率和绩效

减少不必要的营销经费和促销花费

提升消费者对品牌的黏性

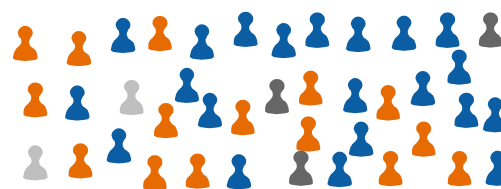
提供细分的相应模型的基础, 提升相应模型精度



优化前



基于分类结果的市场营销

优化后



  = 最优顾客

几种不同商业需求出发的分群类型



需求与态度

依据调查问卷结果针对需求的数据分群



生命周期

依据顾客的消费周期和需求分群



行为特征

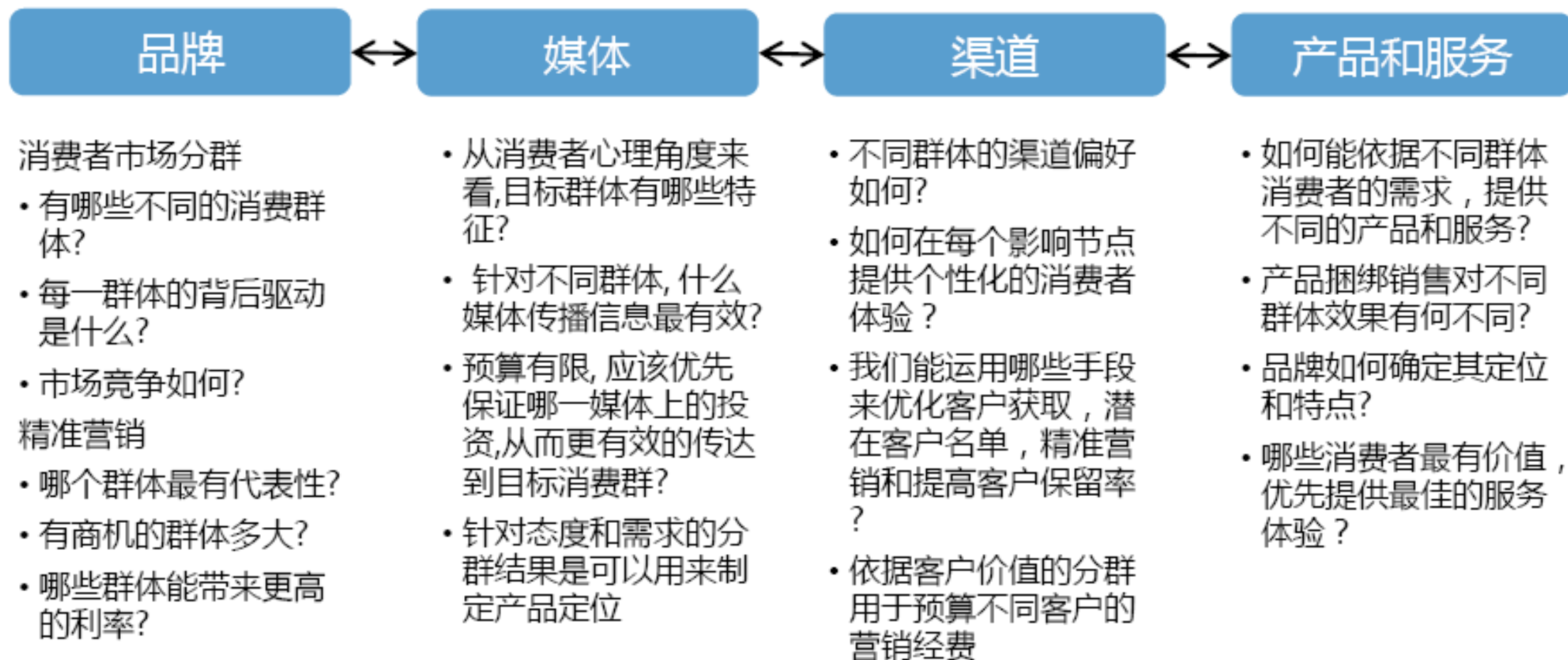
依据消费记录,个人信息分群



客户价值

依据顾客的潜在价值分群

客户分群在商业上的应用



一个电信公司的客户行为特征画像

Group 1	Group 2	Group 3	Group 4	群组	营销 / 战术
喜欢深夜通话 多发短信 多发话给预付费用户 忙时发话少	多发短信 多发话给后付费用户 忙时发话少	高通话量 发话网外多 有国际长途	交往圈固定 多发话给后付费用户	1	夜间通话套餐；短信大礼包；转为预付费；激励话音用量
				2	闲时通话套餐/激励；短信大礼包；激励话音用量
				3	高价值客户大礼包；激励国际长途通话；高端资费方案
				4	高价值客户大礼包；业务引荐奖励
				5	激励话音用量；高价值客户大礼包
				6	移动-固网通话套餐；晨间通话套餐
				7	短信激励；忙时通话套餐
				8	夜间通话套餐；闲时通话套餐；通话量奖励
Group 5	Group 6	Group 7	Group 8		
高通话量 发话网外多	固网通话多 高通话量 喜欢早晨通话	多发短信 多发话给后付费用户 忙时发话多	高通话量 喜欢深夜通话 闲时通话多		

一个银行的多维度客户画像

产品周期 行为模式

投资偏好、消费模式、产品周期

客户结构

客户P&L,利润贡献、产品持有

生命阶段

分析客户所处的生命阶段：学生、年轻夫妇、养育子女等

客户群	占比	客户管理策略	活动示例
理财到期	16.5%	◇成长型培育 ◇主动关怀	◇回报较前次更好的理财或保险产品 ◇吸引他行储蓄转移
基金偏好	4.0%	季节型培育	◇季度性理财产品推荐 ◇或者投资型更强的产品
网购达人	12.6%	◇客户体验提升 ◇成长型培育	◇客户体验提升种子选手 ◇理财指导
房贷车贷	13.1%	◇成长型培育 ◇客户关怀	◇解除抵押提醒 ◇理财指导
涨薪一族	36.8%	◇主动关怀 ◇粘滞型培育	◇信用卡发卡 ◇基金定投，养老定投 ◇理财指导
大额取款	7.7%	◇主动关怀 ◇粘滞型培育	◇信用卡 ◇小额贷款推荐
新添儿女	2.5%	◇成长型培育 ◇电子渠道培育	◇根据儿童成长阶段，推荐教育理财或儿童、养老保险 ◇电子渠道portfolio DIY


客户分群的算法



有明确目标的我们一般会正对目标变量建立相应的模型（响应/预测），用模型评分做排序划分群组。

而无明确目标的客户分群通常采用**聚类分析**的算法，其目标是尽量将相似的研究对象(客户)聚集在同一个类别(群体)，同时让相异的客户分布在不同群体。因此我们需要定性且定量的去描述相似/类似或者相异同的“度”，统计上我们使用“**距离**”。

3、聚类的基本逻辑



基本逻辑

步骤1：从N个观测和K各属性数据开始；

步骤2：计算N各观测两两之间的距离；

Subjects	1	2	3	...	N
1		1.782	2.538	...	47.236
2	1.782		0.821	...	39.902
3	2.538	0.821		...	41.652
...
N	47.236	39.902	41.652	...	

步骤3：将相离最近的观测聚为一类，将距离远的分为不同的类。最终达到组间的距离最大化，组内的距离最小化。

- 层次聚类

形成类相似度层次图谱，便于直观的确定类之间的划分。该方法可以得到较理想的分类，但是难以处理大量样本。

- 非层次聚类（K均值法）

将观测分为预先指定的，不重叠的类。该方法可以处理样本量巨大的数据，但是不能提供类相似度信息，不能交互的决定聚类个数。

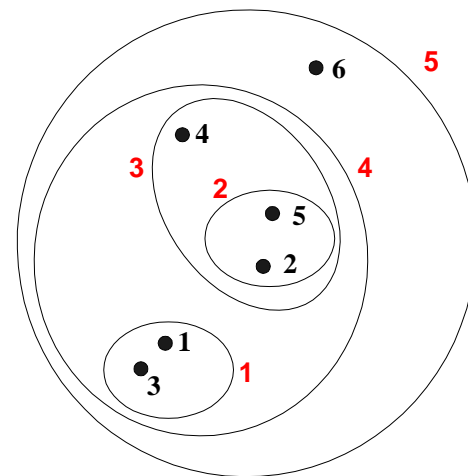
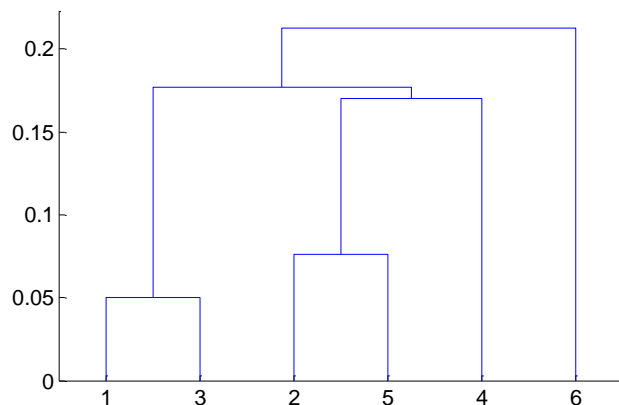
- 两步法聚类

先使用K均值法聚类，然后使用层次方法。

4、系统聚类

层次聚类

- 建立类之间的层次关系
- 通过层次树决定聚类个数和聚类方式



•基本步骤：

1. 计算每两个观测之间的距离
2. 将最近的两个观测聚为一类，将其看作一个整体计算与其它观测（类）之间的距离
3. 一直重复上述过程，直至所有的观测被聚为一类

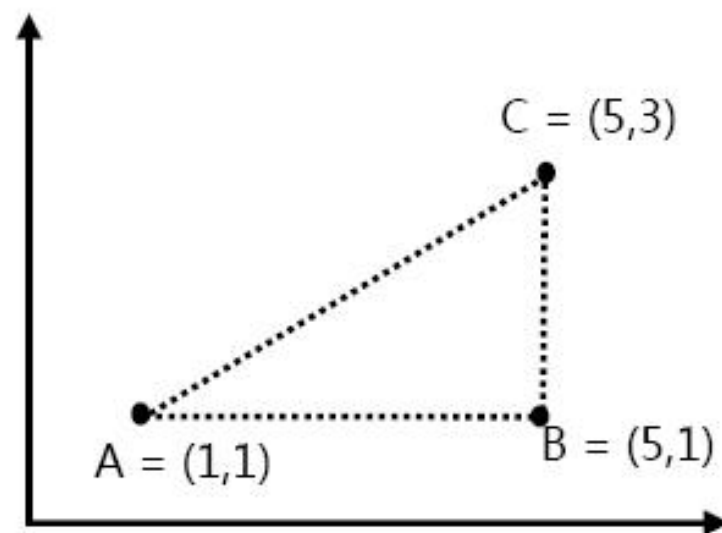
定义两个观测(两点)之间的距离

首先我们来认识一下距离：

两个观测，其属性我们用一个向量来表达

第*i*个观测值 $X_i \sim (x_{i1}, x_{i2}, \dots, x_{ip})$

第*j*个观测值 $X_j \sim (x_{j1}, x_{j2}, \dots, x_{jp})$



➤ 绝对值距离

$$d_{ij}(1) = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

➤ 欧氏距离

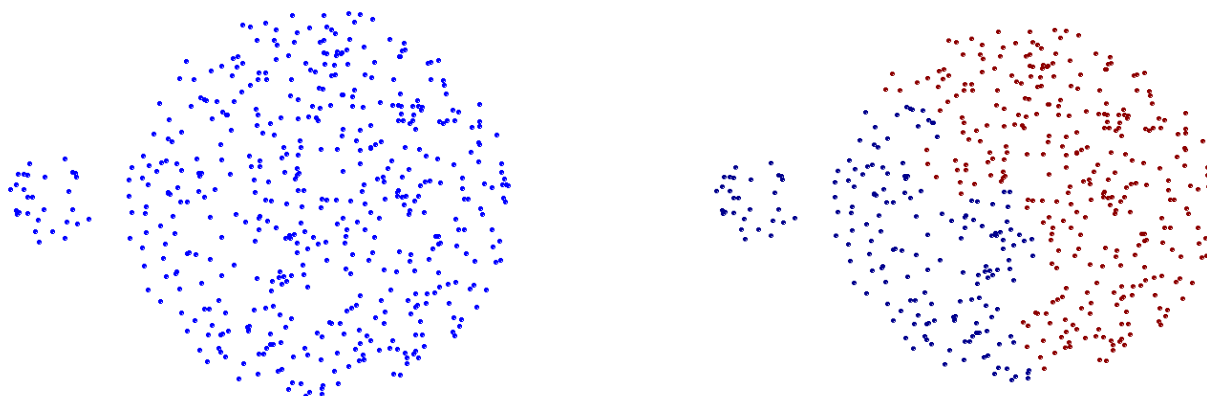
$$d_{ij}(2) = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{1/2}$$

➤ 明考斯基(Minkowski) 距离 $d_{ij}(q) = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^q \right]^{1/q}$



定义两个类(两椭圆球)之间的距离

平均联接:



- 倾向于将大的类分开
- 所有的类倾向于具有同样的直径
- 对异常值敏感

定义两个类(两椭圆)之间的距离

重心法:两个类各自重心之间的距离。

$$D_{centroids}(C_i, C_j) = d(r_i, r_j)$$

- 较少受到异常值的影响，但因为群间的距离没有单调递增的趋势，在树状聚类图上可能出现图形逆转，限制了它的使用。

Ward最小方差法:各个观测之间的利差平方和最小。

$$D_w(C_i, C_j) = \sum_{x \in C_{ij}} (x - r_{ij})^2 - \sum_{x \in C_i} (x - r_i)^2 - \sum_{x \in C_j} (x - r_j)^2$$

- 较少受到异常值的影响，适用范围广。

Ward最小方差法

$$SS = \sum_{i=1}^p \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2$$

	X	Y
A	6	5
B	7	6
C	2	4
D	4	2
E	2	1

	A	B	C	D	E
A					
B	2				
C	17	29			
D	13	25	8		
E	32	50	9	5	

以前述的資料為例，一開始如果將 AB 合併成一集群，則此時 $SS = (6 - \frac{6+7}{2})^2 + (7 - \frac{6+7}{2})^2 + (5 - \frac{5+6}{2})^2 + (6 - \frac{5+6}{2})^2 = 1$ ，如果將 CD 合併成一集群，則 $SS = (2 - \frac{2+4}{2})^2 + (4 - \frac{2+4}{2})^2 + (4 - \frac{4+2}{2})^2 + (2 - \frac{4+2}{2})^2 = 4$ 。

當合併成三個集群時，如果是 AB、CD，及 D 的組合，則聯合組內 SS 就等於 $1+4+0=5$ 。

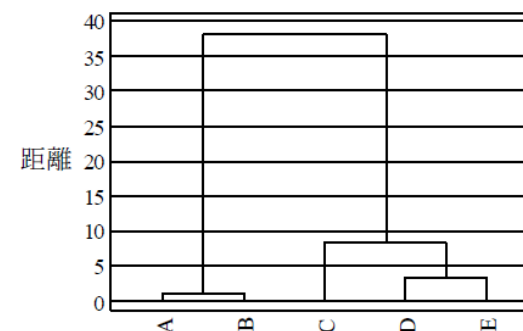
Ward最小方差法

$$SS = \sum_{i=1}^p \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2$$

	集群				聯合組內 SS
	1	2	3	4	
4 個集群之可能組合					
1	AB	C	D	E	1.00
2	AC	B	D	E	8.50
3	AD	B	C	E	6.50
4	AE	B	C	D	16.00
5	BC	A	D	E	14.50
6	BD	A	C	E	12.50
7	BE	A	C	D	25.00
8	CD	A	B	E	4.00
9	CE	A	B	D	4.50
10	DE	A	B	C	2.50

Ward最小方差法

華德法



3 個集群之可能組合					
1	ABC	D	E		16.00
2	ABD	C	E		13.33
3	ABE	C	D		28.00
4	AB	CD	E		5.00
5	AB	CE	D		5.50
6	AB	DE	C		3.50
2 個集群之可能組合					
1	ABC	DE			18.50
2	AB	CDE			8.33
1 個集群之可能組合					
1	ABCDE				38.00

要点1：要预先处理变量

- 收到的数据通常需要经过处理才能用于分析：
 - 缺失值
 - 异常值（极大或极小）
 - 分类变量需要转化为哑变量（0/1数值）
 - 分类变量类别过多
- 不同的统计方法对数据有不同的要求：
 - 决策树允许缺失值和异常值
 - 聚类分析和回归模型则不支持缺失值

要点2：变量标准化

为什么要做标准化：

变量的量纲的不一样引起计算距离的偏差

比如我们用了两个维度：收入和年龄

收入的取值范围 [\$10,000, \$100,000]

年龄的取值范围 [18, 100]

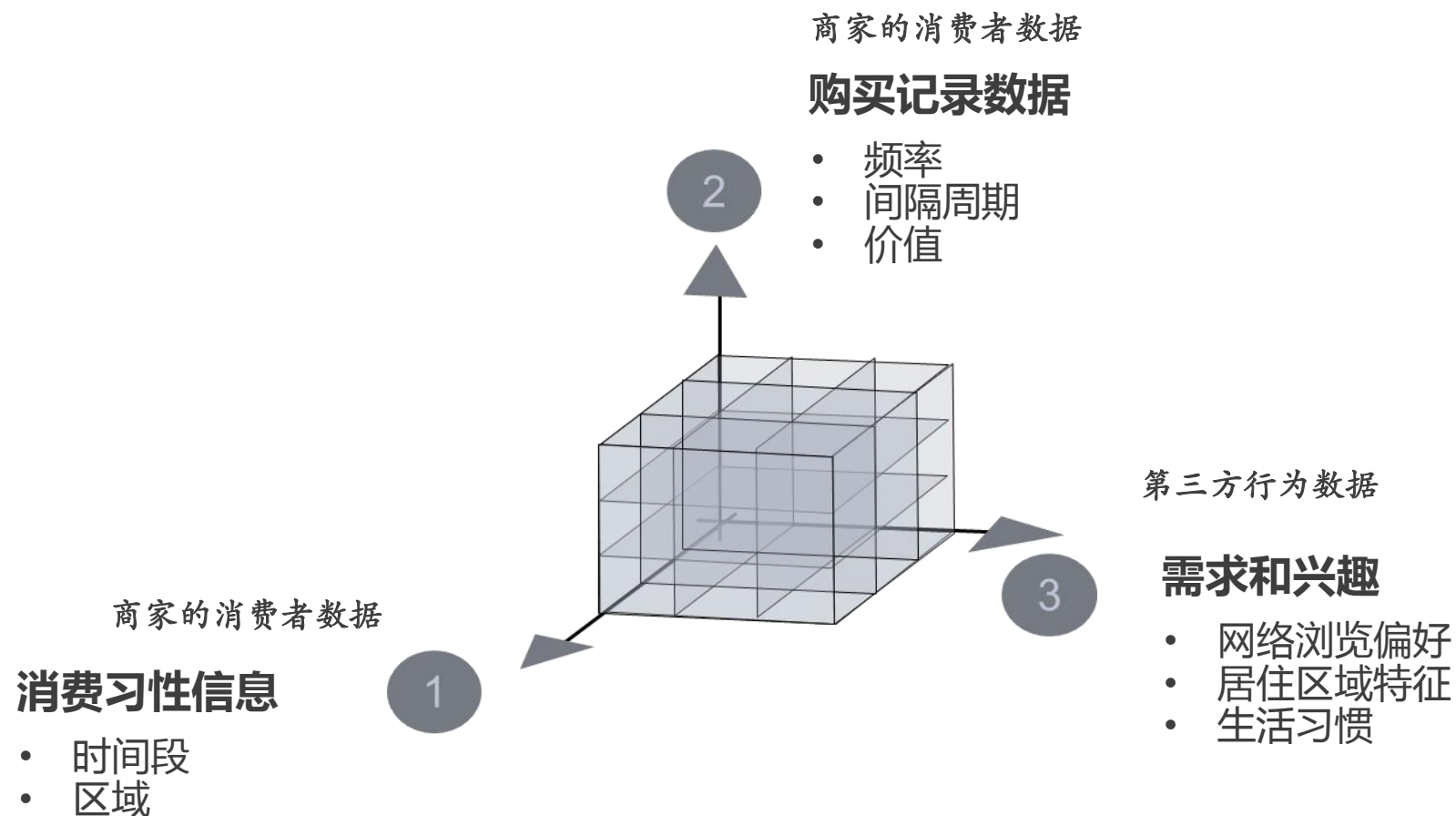
方法一： 中心化

$$std(x_{ip}) = \frac{x_{ip} - \bar{x}_p}{S_p}$$

方法二： 极差标准化

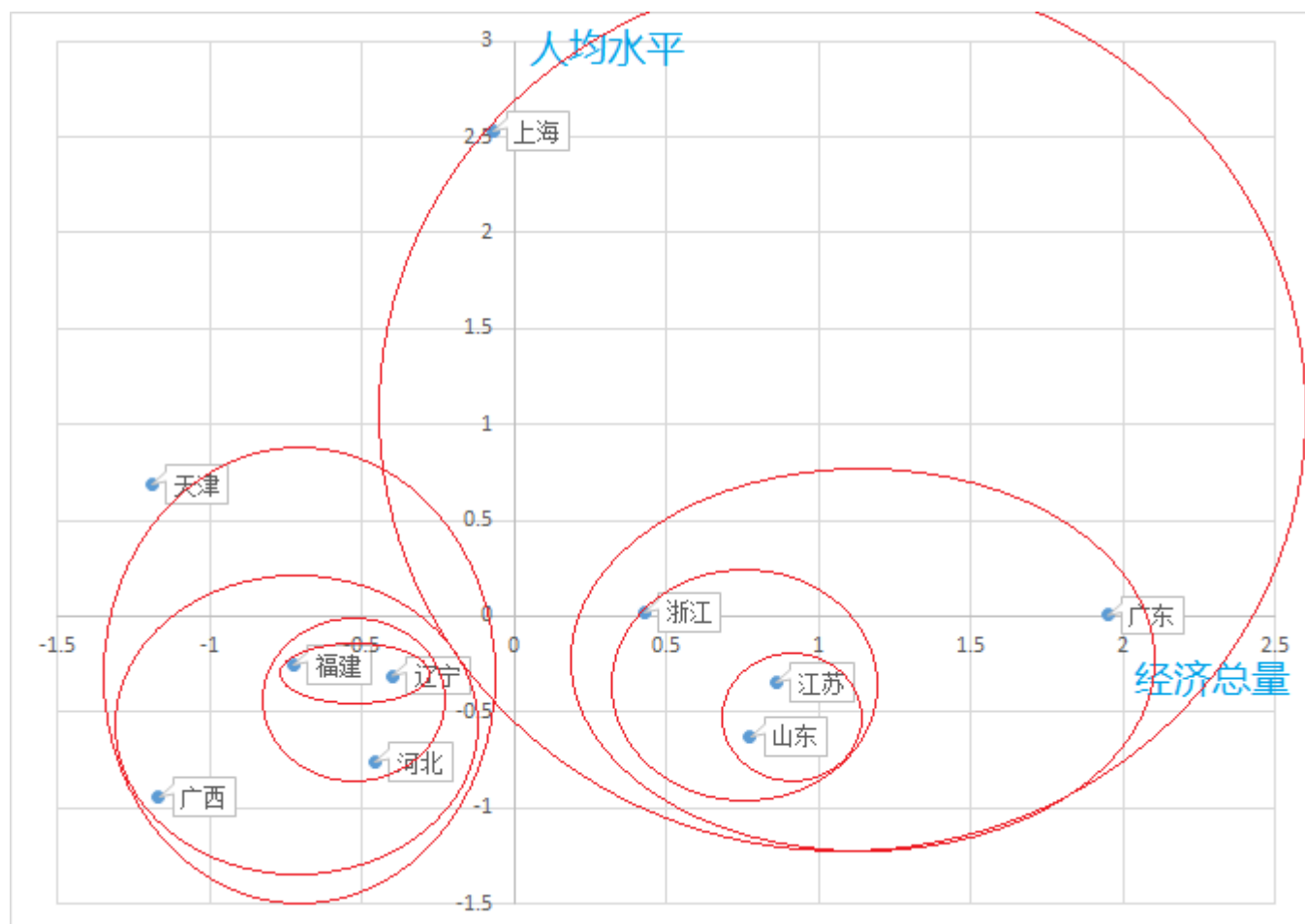
$$x - \min(x) / \max(x) - \min(x)$$

要点3：不同维度的变量,相关性尽量低



使用 “CITIES_10” 演示系统聚类法。注意因子分析。

由于本次聚类只使用了两个因子，因此可以在二维平面上展现聚类的过程。

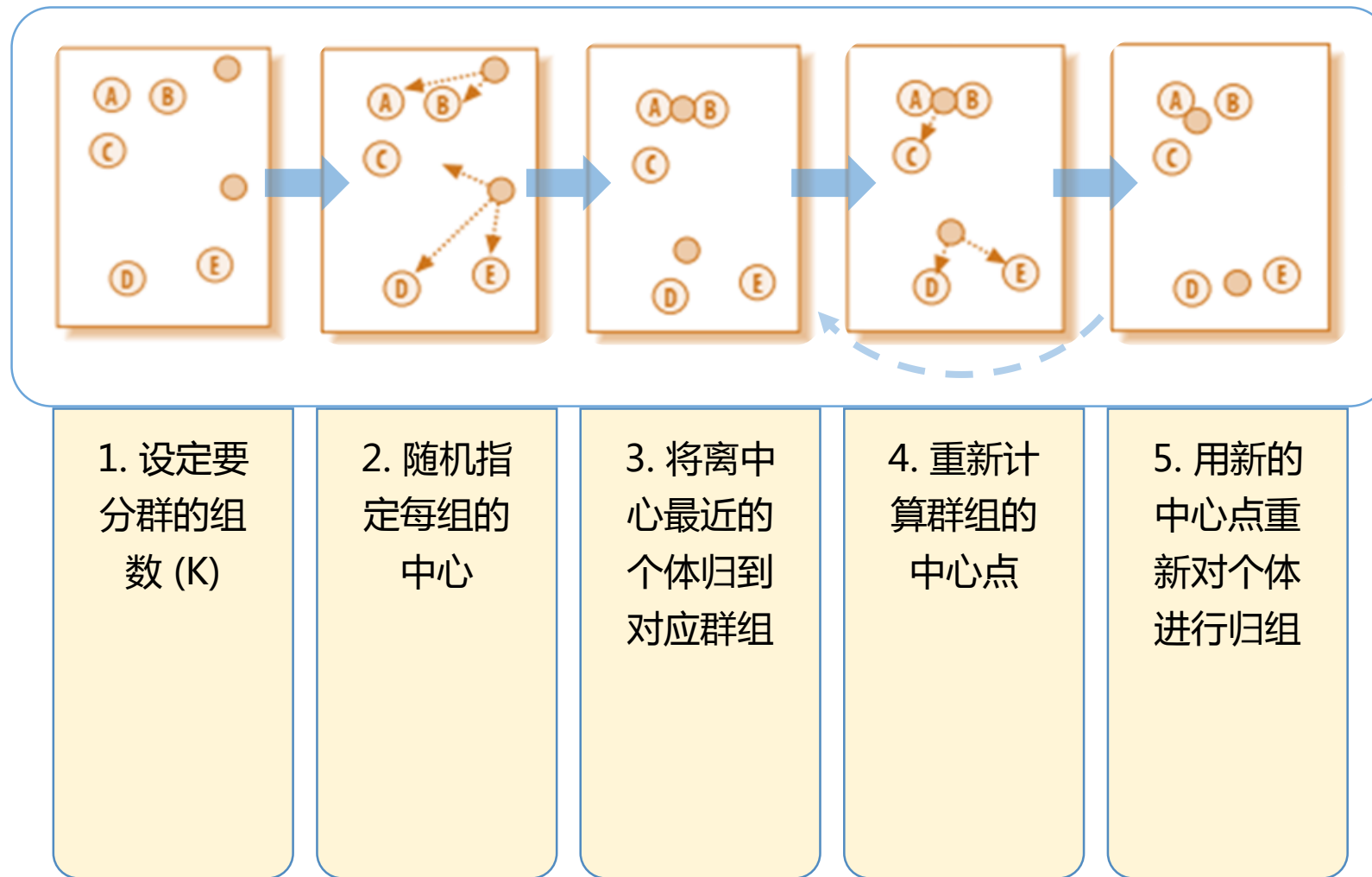


5、K-means聚类

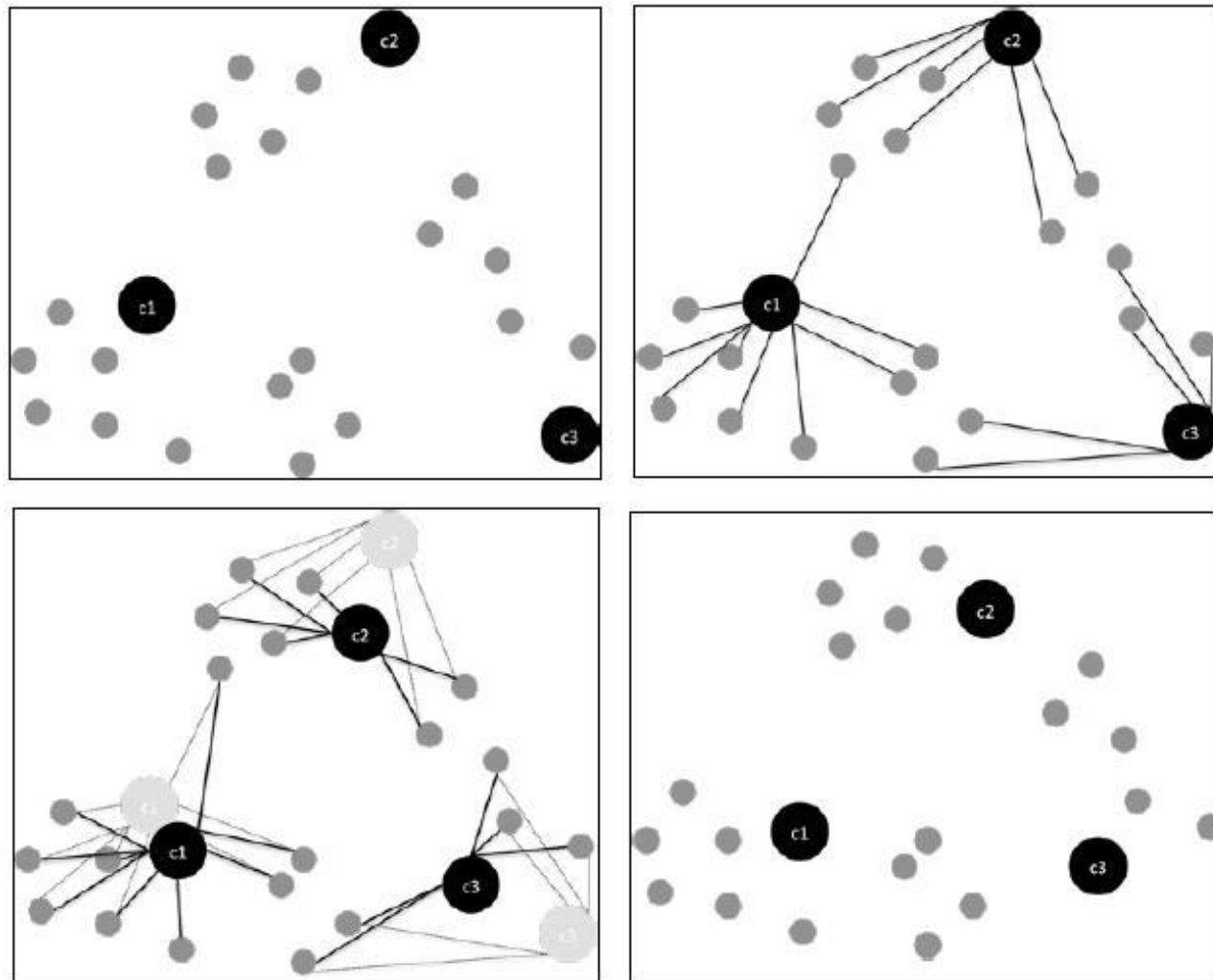
K-means聚类过程

- 设定K值，确定聚类数（软件随机分配聚类中心所需的种子）；
- 计算每个记录到类中心的距离（欧式），并分成K类。
- 然后把K类中心（均值），作为新的中心，重新计算距离；
- 迭代到收敛标准停止（最小二乘准则）。

K-Means 聚类方法步骤



K-Means 聚类方法步骤



要点1：预先处理变量的缺失值、异常值

要点2：变量标准化

要点3：不同维度的变量,相关性尽量低

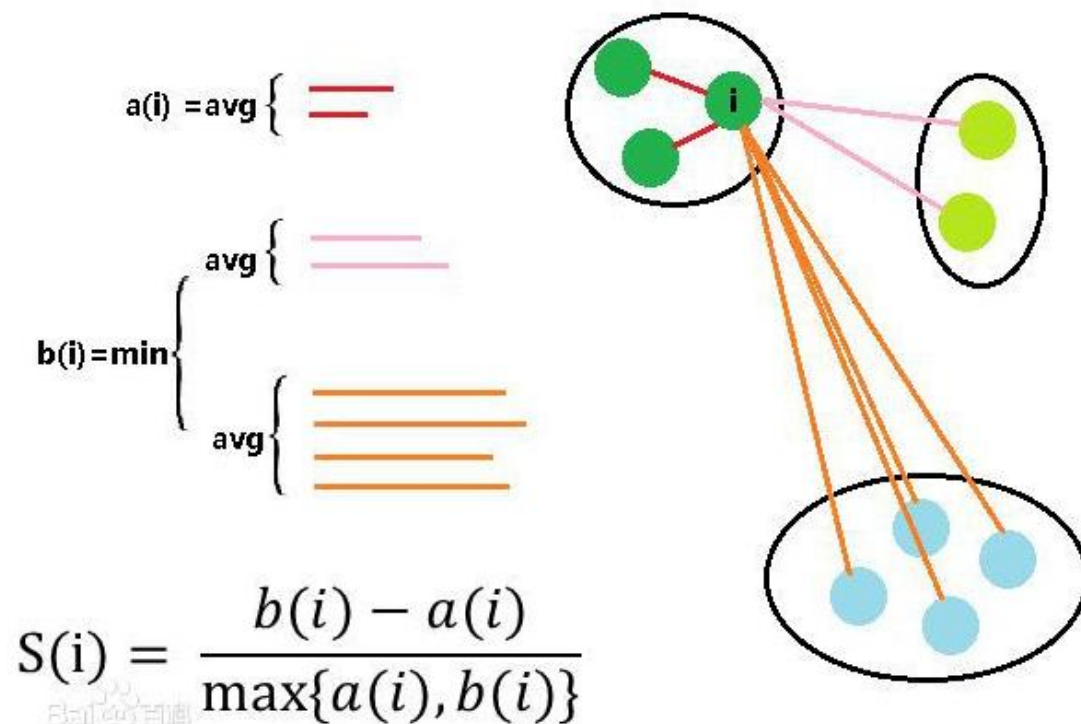
要点4：如何决定合适的分群个数？

- 主要推荐轮廓系数(Silhouette Coefficient)，并结合以下注意事项：
 - 分群结果的稳定性
 - 重复多次分群,看结果是否稳定
 - 分群结果是否有好解释的商业意义

轮廓系数(Silhouette Coefficient)

轮廓系数可以认为是WARD法在K-means聚类中的一种拓展，其思路如下：

- 1、我们虽然不知道应该正确的讲样本划分为几类，但是我们知道K-means法的分类是最好的；
- 2、最好的分类结果应该是组间变异（方差）最大、组内变异最小的那个分组；
- 3、遍历所有的分组数量，找到使2最优的那个分组数量。



$a(i)$: i 向量到同一簇内其他点不相似程度的平均值

$b(i)$: i 向量到其他簇的平均不相似程度的最小值

轮廓系数的值是介于 $[-1, 1]$ ，越趋近于1代表内聚性和分离度越好。

快速聚类的两种运用场景

1、发现异常情况：如果不对数据进行任何形式的转换，只是经过中心标准化或级差标准化就进行快速聚类，会根据数据分布特征得到聚类结果。这种聚类会将极端数据聚为几类。方法一会演示这种情况。这种方法适用于统计分析之前的异常值剔除，对异常行为的挖掘，比如监控银行账户是否有洗钱行为、监控POS机是否有从事套现、监控某个终端是否是电话卡养卡客户等等。

2、将个案数据做划分：出于客户细分目的的聚类分析一般希望聚类结果大致平均的几大类，因此需要将数据进行转换，比如使用原始变量的百分位秩、Turkey正态评分、对数转换等等。在这类分析中数据的具体数值并没有太多的意义，重要的是相对位置。方法二会演示这种情况。这种方法适用场景包括客户消费行为聚类、客户积分使用行为聚类等等。

- 使用 PROFILE_BANK 数据演示K均值法。

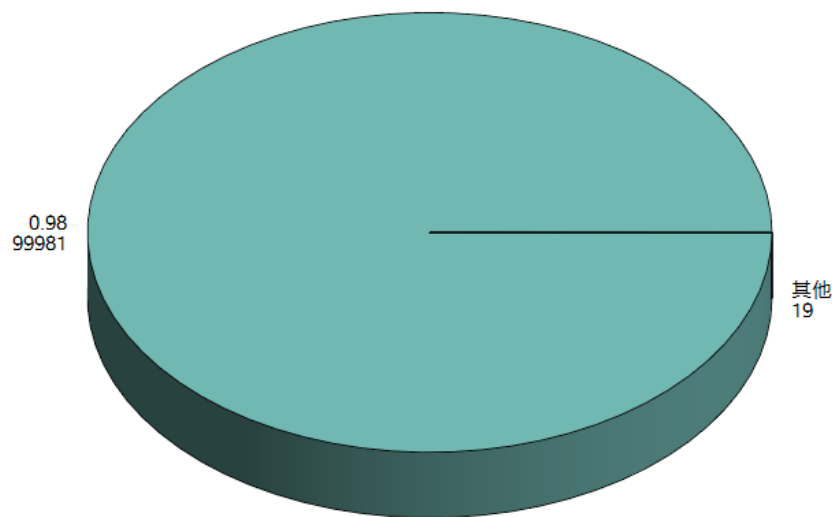
方法一：发现异常个体 不进行分布形式转换

首先分析每个变量的分布情况，发现严重的偏态分布，但是并不将变量转化为正态分布,而是用聚类算法默认的学生标准化。

方法一：结果展现

使用饼形图对聚类模型的聚类结果作描述，得到如下所示饼形图。

饼形图反映绝大部分客户被分为一类，其它三类的样本量非常少。



这种情况在聚类中非常多见，主要的原因是输入变量严重右偏。这是变量分布自然的反映。因此聚类算法可以用于异常值检验和反常情况侦测，不如洗钱。但是出于营销或客户维护角度的客户画像任务要求将客户均匀的分若干类。

方法二：客户分群 进行正态分布转换

变量归一化v.s.分布转换

在进行多元统计分析是，进行适当的变量归一化是得到优良结果的前提条件，部分多元统计方法（主成分、因子分析、聚类等）提供默认的归一化方法，但不能满足数据分布多样化的需求。

变量转换有两种：

消除量纲但是不改变分布（归一化）：中心标准化、极差标准化；
同时消除量纲与改变分布：对数、百分位秩、Tukey评分等。

以下使用Profile_telecom数据对其中的cnt_call（通话次数）变量进行归一化。

中心标准化

该归一化方法之前章节讲过，是主成分等分析方法默认的，因此用处不大。

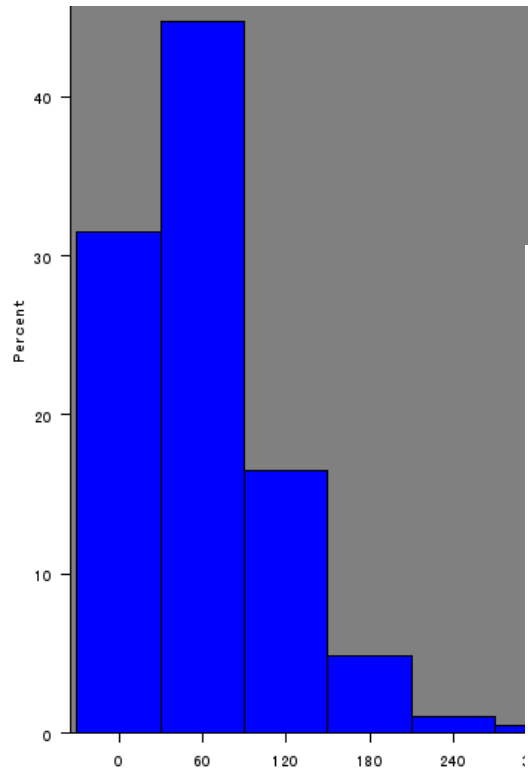
$$A = \frac{x_i - \text{mean}(x)}{\text{std}(x)}$$

极差标准化

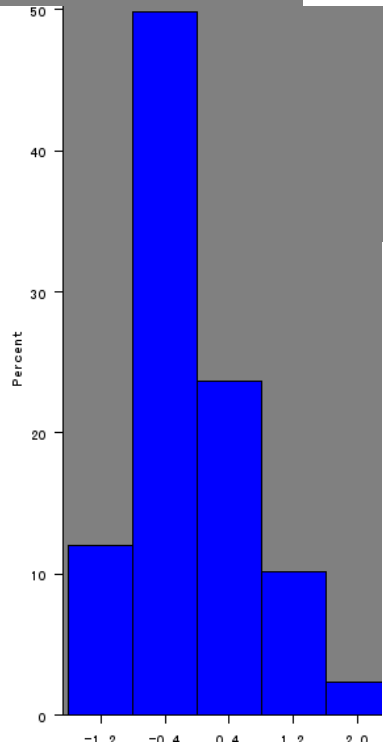
该方法和中心标准化类似，只不过值域为[0,1],该方法没有命令可以完成，只能根据公式编程完成。

$$A = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

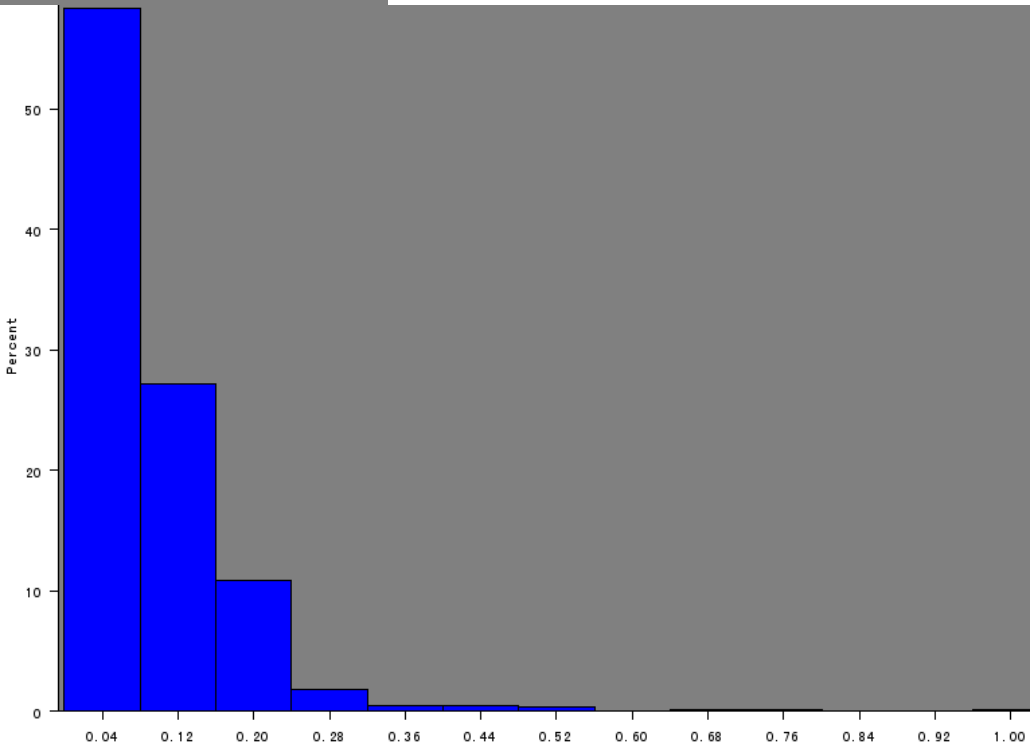
原始变量



中心标准化



极差标准化



百分位秩

变量从小到大排序，然后依次赋予序列号，最后用总的样本量除以序列号，值域[0,100]。

Tukey正态分布打分

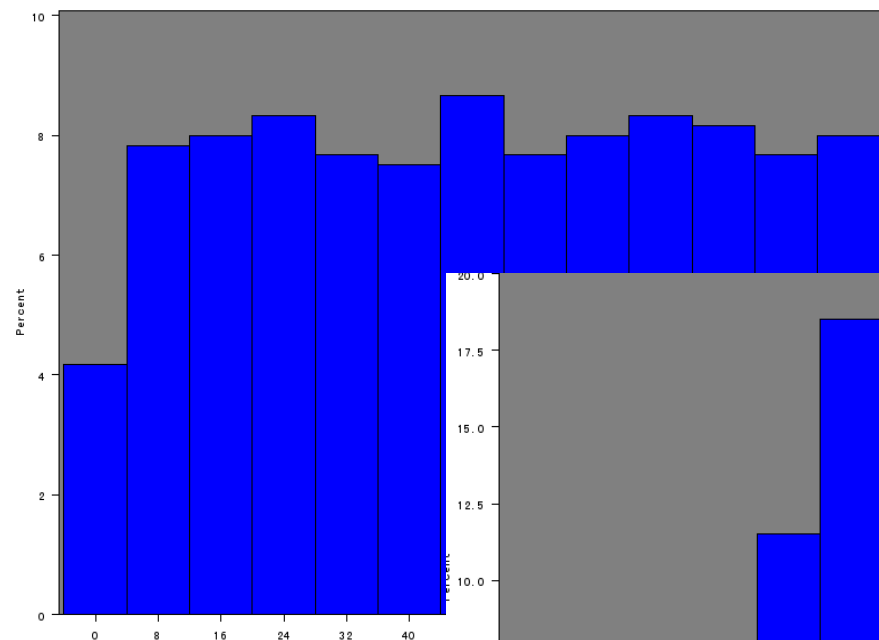
先转化为百分位秩，然后转化为正态分布。

变量取自然对数

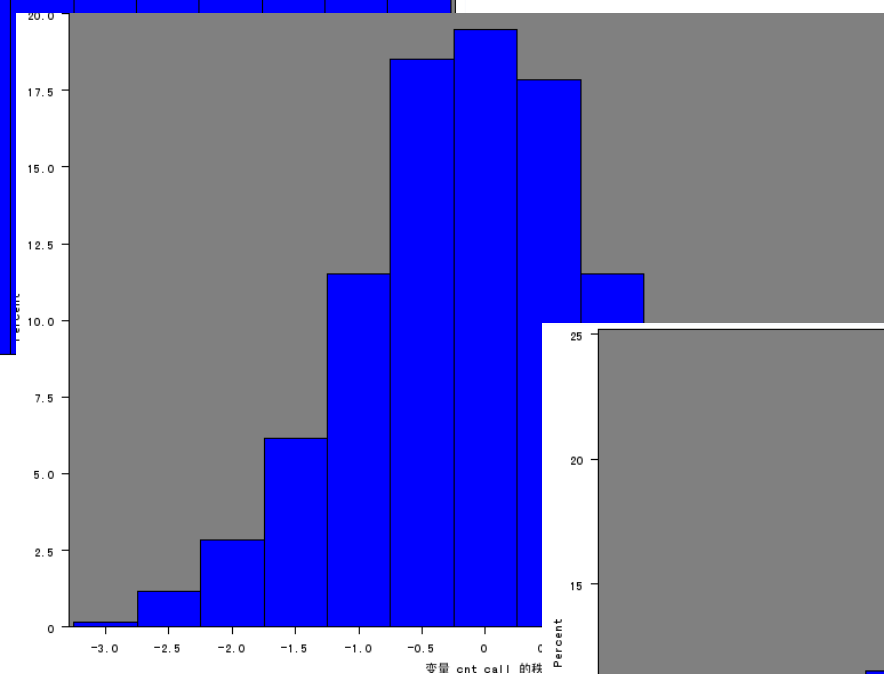
数学表达式：

$$A = \ln(x)$$

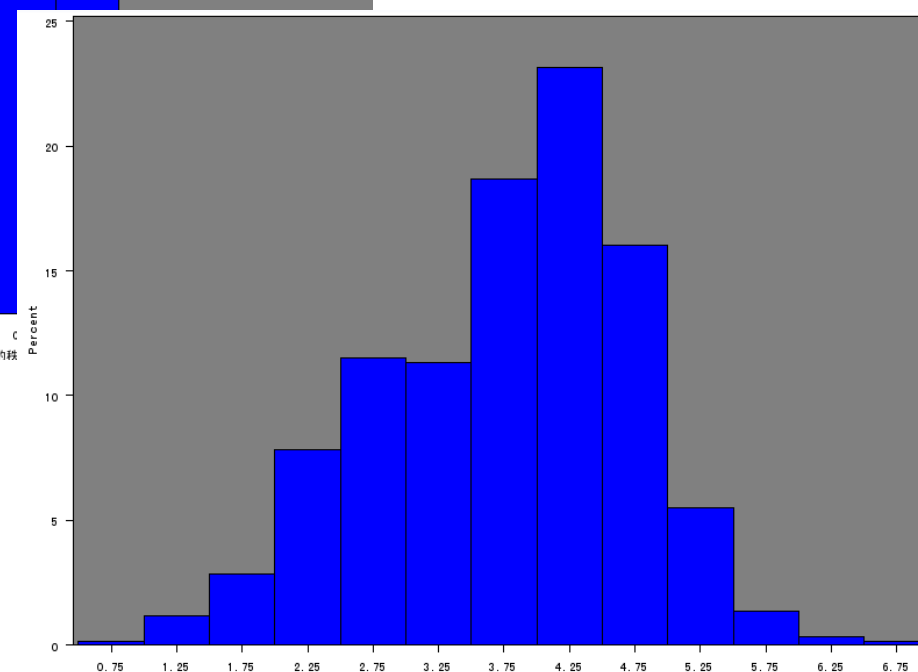
百分位秩



Tukey正态分布打分



变量取自然对数

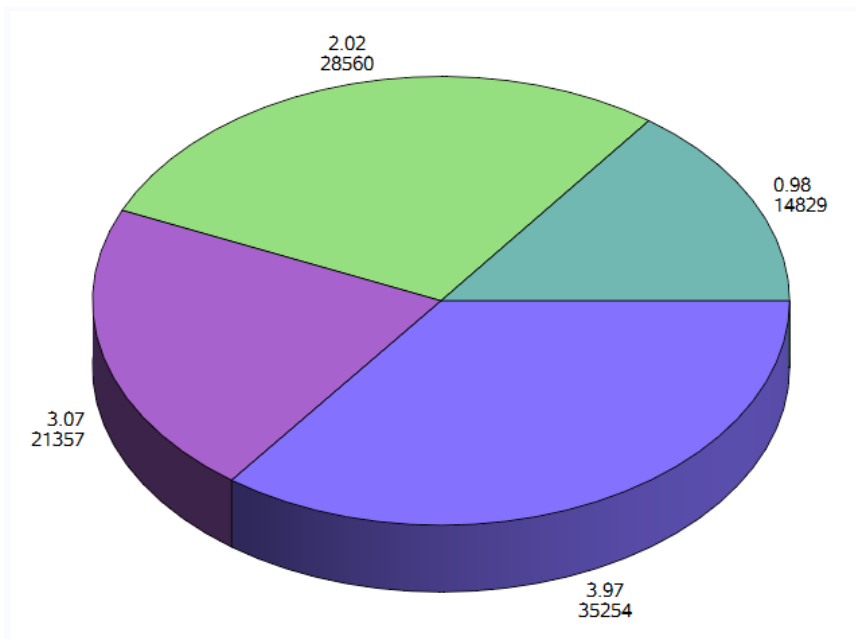


变量转换总结

非对称变量在**聚类分析**中选用**百分位秩**和**Tukey正态分布****打分**比较多；
在**回归分析**中**取对数**比较多。因为商业上的聚类模型关心的客户的排序情况，回归模型关心的是其具有经济学意义，对数表达的是百分比的变化。

方法二：结果展现

再次得到如下所示饼形图，这次的聚类结果中每个类的样本量大体一致。



饼形图反映绝大部分客户被分为一类，其它三类的样本量非常少。

这种情况在聚类中非常多见，主要的原因是输入变量严重右偏。这是变量分布自然的反映。因此聚类算法可以用于异常值检验和反常情况侦测，不如洗钱。但是出于营销或客户维护角度的客户画像任务要求将客户均匀的氛围若干类。

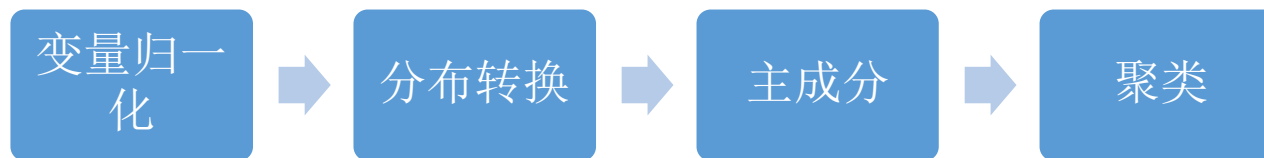
方法二：结果展现

通过描述统计，为每个类取一个便于记忆的名字。

聚类	观测的个数	变量	均值	N	中位数	类别	表现	猜测	名称
1	53555	Factor1	-0.2997788	53555	-0.3461975	1	两方面的渠道都是中下,有偿服务居中。	有一定继续的中年人，新兴渠道和柜台渠道均衡发展。	一般中年客户
		Factor2	-0.6084497	53555	-0.5631022				
		Factor3	0.0381577	53555	0.0526953				
2	17081	Factor1	-0.6659119	17081	-1.0414812	2	新兴渠道较少，柜台服务最高，有偿服务较少。	手中资金不太充沛的老年客户。	中价值老年群体
		Factor2	1.2225384	17081	1.1212861				
		Factor3	-0.5457501	17081	-0.9162650				
3	23323	Factor1	1.3488048	23323	1.2774422	3	新型渠道使用最高，传统渠道中上，但是有偿业务使用中下。	较高的渠道使用有可能是交易需要，手中资金不够充裕。	年青潜在客户
		Factor2	0.2903100	23323	0.2836008				
		Factor3	-0.1554556	23323	-0.0940825				
4	6041	Factor1	-0.6669555	6041	-0.8298383	4	新兴渠道使用最少；传统渠道和有偿服务使用最多。	手中资金充沛的中老年客户，业务很多，但是对新渠道不够熟悉	高价值老年客户
		Factor2	0.8164950	6041	0.7830382				
		Factor3	1.8050178	6041	1.7089357				

流程：

一般情景下的聚类：



发现异常情况的聚类：

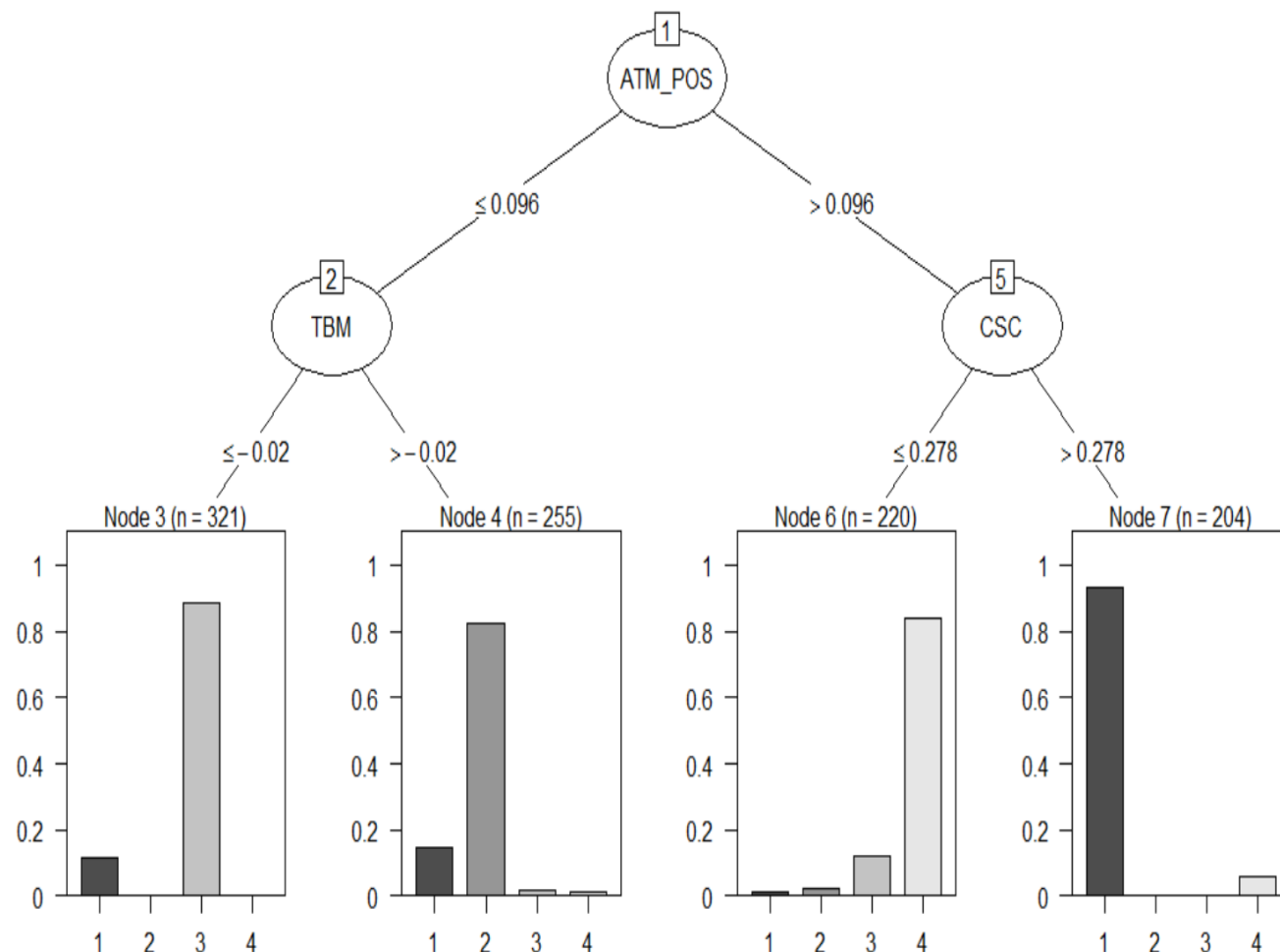


总结：

- 聚类模型对分析人员的业务修养要求较高；
- 聚类结果好坏不是看统计指标就可以有明确的答案的。统计指标是在所有的变量都符合某个假设条件才能表现良好的，而实际建模中很少有能达到那种状态；
- 聚类的结果要详细的作描述性统计，甚至作抽样的客户访谈，以了解客户的真实情况，所以让业务人员满足客户管理的目标，是聚类的终极目标。

6、使用决策树做聚类后客户分析

决策树的分组画像和规则



```
Rule 1: (204/14, lift 3.5)
ATM_POS > 0.09552076
CSC > 0.2783226
-> class 1 [0.927]

Rule 2: (254/44, lift 3.8)
ATM_POS <= 0.09552076
TBM > -0.02008908
-> class 2 [0.824]

Rule 3: (322/37, lift 2.8)
ATM_POS <= 0.09552076
TBM <= -0.02008908
-> class 3 [0.883]

Rule 4: (220/35, lift 4.2)
ATM_POS > 0.09552076
CSC <= 0.2783226
-> class 4 [0.838]
```

秦路主讲

七周成为数据分析师

七周为期，Get一条数据分析师职业黄金通道！



Python

数据分析与挖掘

集Python爬虫、数据采集、数据处理、数据分析与数据挖掘于一体，打造Python全栈工程师

主讲老师: 韦玮

VIP会员群+在线答疑+录播复习+1年反复观看

参团课程

案例为师,实战为王

开启Python机器学习之路

科学规划全套课程体系,从入门到进阶,从理论到技巧,嵌入丰富课程案例讲解,逐步推进

讲师: 唐宇迪 深度学习领域多年一线实践研究专家

独一无二的 数据仓库建模指南系列教程升级版

- 从企业视角进行数据规划以及数据仓库模型的搭建
- 高质量的数据库模型和技巧,以及丰富的例子
- 数据仓库架构理论和实践要领

资深讲师: BAO胖子 15年+BI从业经验
涉足电力、快消品、医药、信息服务行业的BI老兵

业务知识一站通

技术+业务,挣钱有门路!

讲师: 陈文



自己动手 丰衣足食

Python3网络爬虫实战案例

一循序渐进,案例为王,诠释全面,思路制胜一

讲师: 崔庆才 北航硕士,百万级热度爬文博主



讲师 丘祐玮

人人都爱数据科学家

Python数据科学精华实战课程

数据分析报告制作

秘籍升级版

讲师: 陈丹奕 知乎大神,前百度资深数据分析师

先机致胜 破冰AI

深度学习模型/框架与实战

讲师: 唐宇迪 同济大学硕士
深度学习领域多年一线实践研究专家



BI、商业智能
数据挖掘 大数据
数据分析师
R语言 Python
机器学习
深度学习
人工智能
Hive Hadoop
Tableau
BIEE ETL
数据科学家
PowerBI