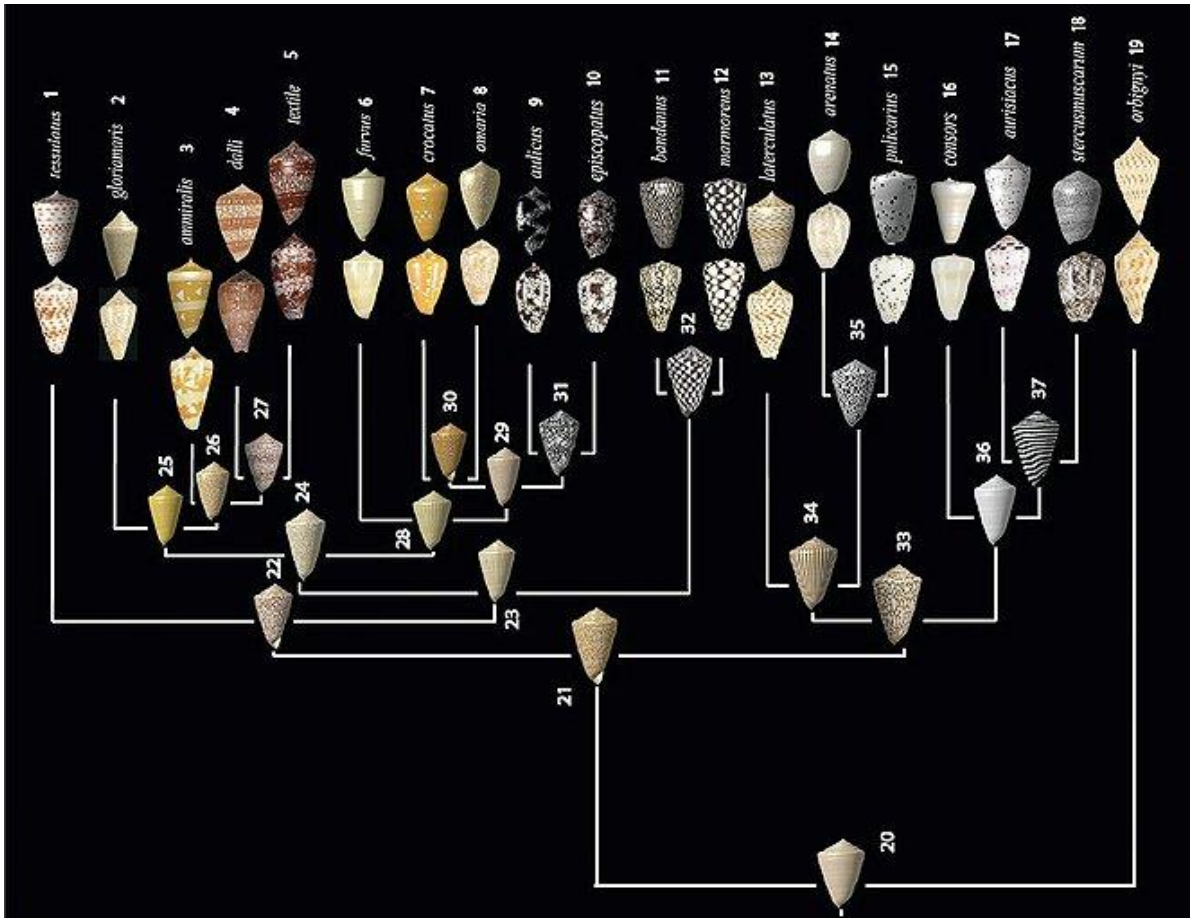


机器学习 第10周

DATAGURU专业数据分析社区

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

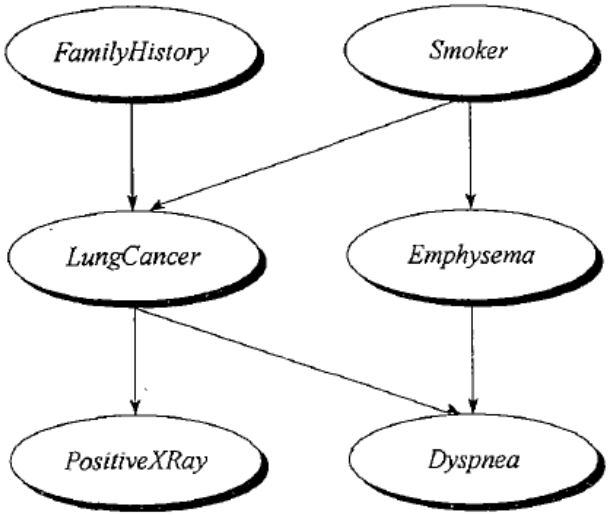
贝叶斯信念网络

- Bayes Belief Network，简称BBN
- 朴素贝叶斯分类器需要特征之间互相独立的强条件，制约了模型的适用
- 用有向无环图表达变量之间的依赖关系，变量用节点表示，依赖关系用边表示
- 祖先，父母和后代节点。贝叶斯网络中的一个节点，如果它的父母节点已知，则它条件独立于它的所有非后代节点
- 每个节点附带一个条件概率表（CPT），表示该节点和父母节点的联系概率

例子

- 韩家炜书第256页

$$P(LungCancer = yes \mid FamilyHistory = yes, Smoker = yes) = 0.8$$
$$P(LungCancer = no \mid FamilyHistory = no, Smoker = no) = 0.9$$



a)

	FH,S	*FH,~S*	*~FH,S*	*~FH,~S*
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

b)

图 9.1 一个简单的贝叶斯信念网络：a）一个提议的因果模型，用有向无环图表示；b）变量 *LungCance*(*LC*) 的条件概率表，给出其双亲节点 *FamilyHistory* 和 *Smoke* 的每个可能值组合的条件概率。取自 Russell、Binder、Koller 和 Kanazawa[RBKK95]

贝叶斯信念网络

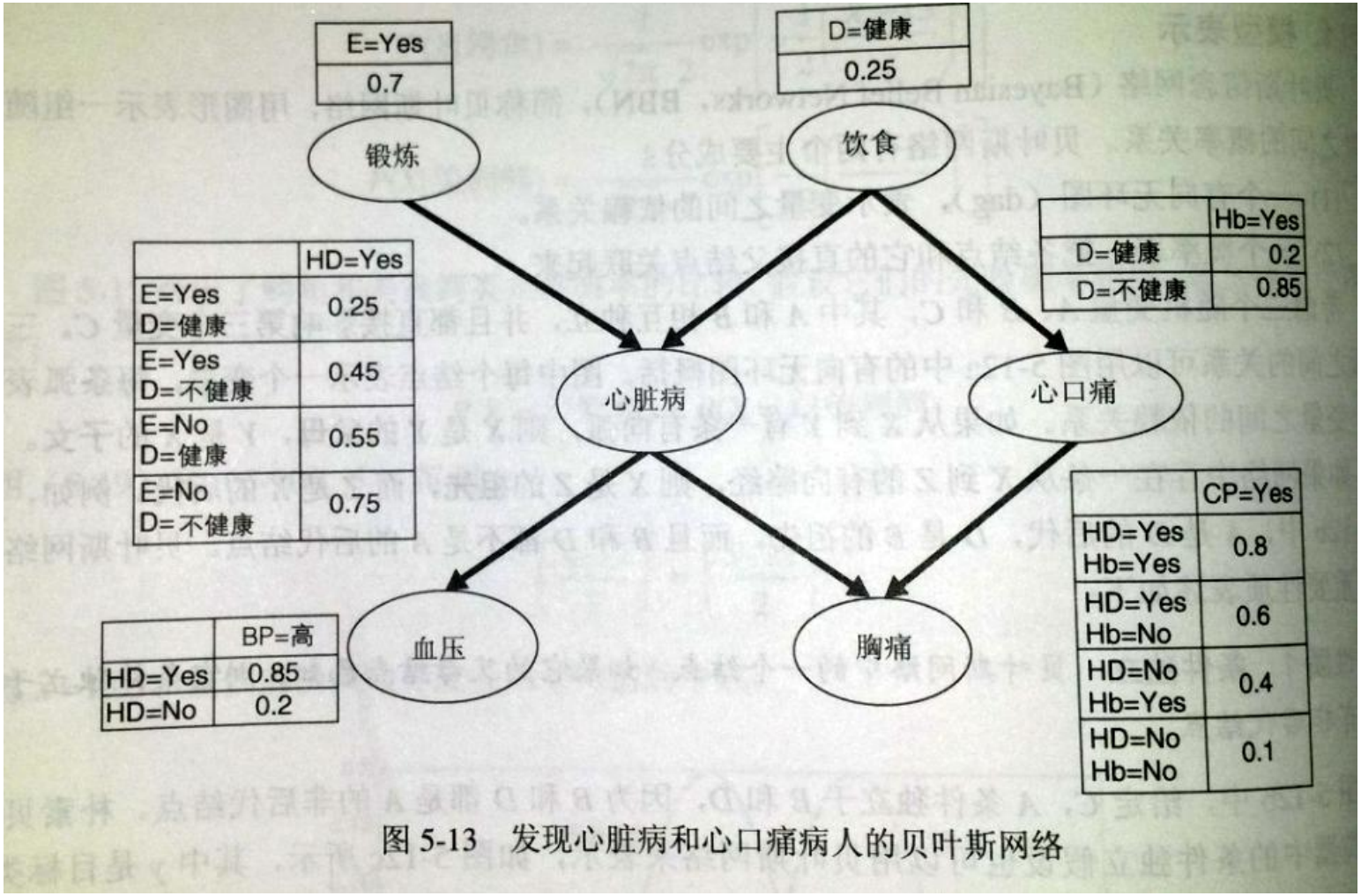


图 5-13　发现心脏病和心口痛病人的贝叶斯网络

训练贝叶斯网络

- 什么时候要训练？存在隐藏变量或隐藏数据
- 转化为类似神经网络求权值的问题，使用梯度下降法进行训练
- 韩家炜书第257页
- 目标函数及解释

$$P_w(D) = \prod_{d=1}^{|D|} P_w(X_d)$$

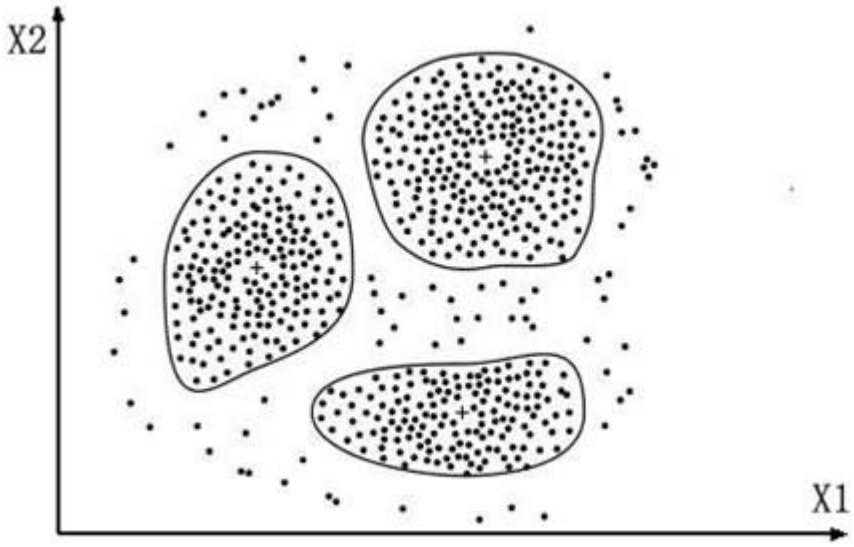
梯度计算及权重更新

$$\frac{\partial \ln P_w(D)}{\partial w_{ijk}} = \sum_{d=1}^{|D|} \frac{P(Y_i = y_{ij}, U_i = u_{ik} \mid X_d)}{w_{ijk}}$$

$$w_{ijk} \leftarrow w_{ijk} + (l) \frac{\partial \ln P_w(D)}{\partial w_{ijk}}$$

聚类

聚类和分类判别有什么区别？



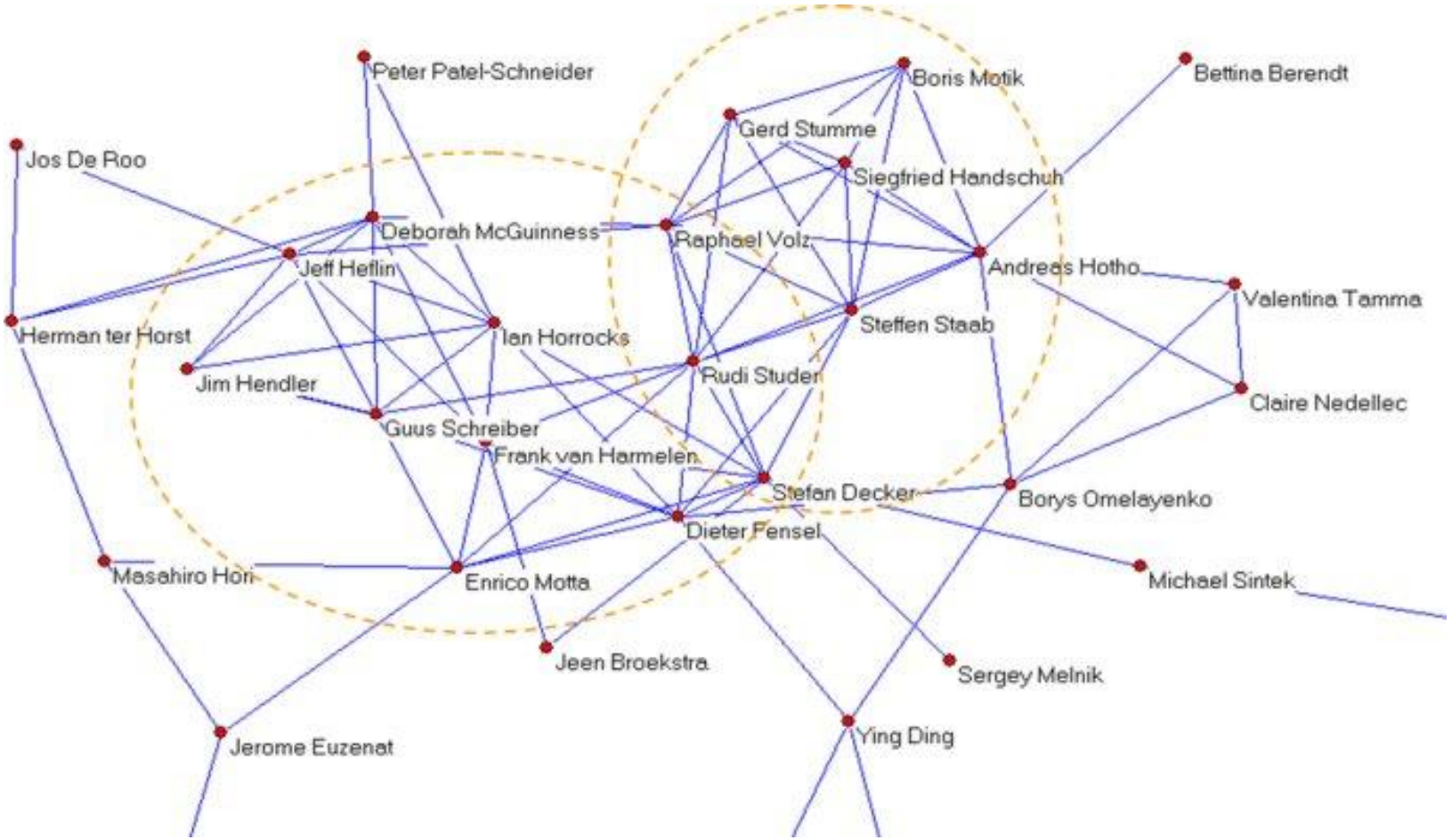
聚类应用场景：寻找优质客户

- 二八定律无处不在
- 20%的用户提供了银行80%的利润来源
- 20%的用户消费了运营商话费总额的80%
- 公司中20%的员工完成了80%的工作
- 社会中20%的人拥有80%的话语权

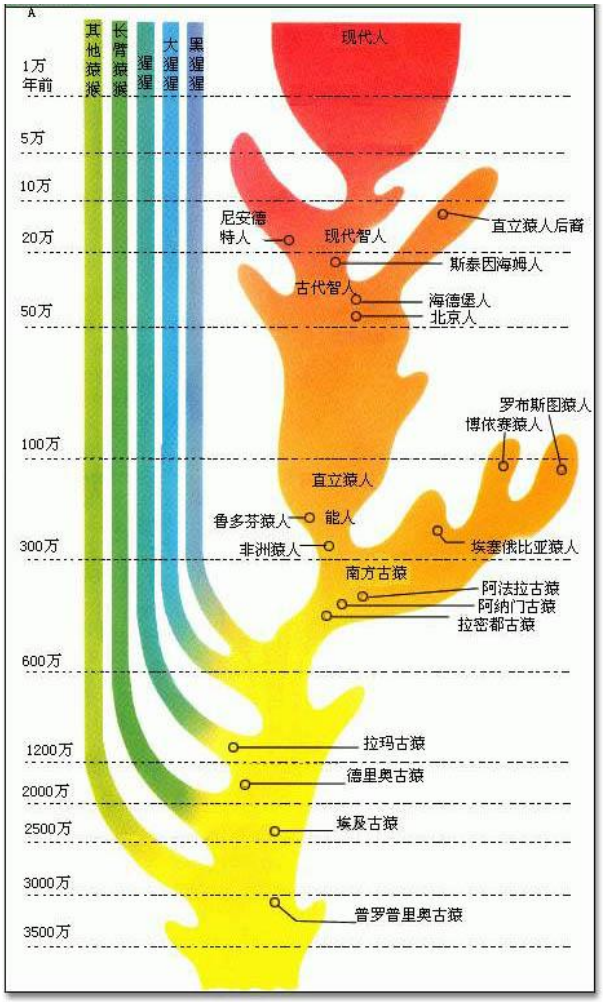
聚类应用场景：推荐系统


图书 > 计算机/网络 > 软件工程/开发项目管理 > 商品详情
看过本商品的还看了
¥137.00
Logitech/罗技 无线鼠标M545_深沉黑_激光级技术
(52条评论)
¥65.00
【当当自营】logitech罗技 M185 无线鼠标 灰色
(931条评论)
¥40.70
大规模分布式存储系统：原理解析与架构实战（阿
(870条评论)
推荐系统
Recommender Systems
An Introduction
分享到：
送积分 472
查看大图
批量购买入口>>
推荐系统(推荐系统必读经典，百度技术委员会主席廖若雪、新浪微博数据挖掘技术专家张俊林、人民搜索商务部总监常兴龙、百分点信息科技有限公司首席运营官张韶峰联袂推荐！)
当当价 ¥47.20 （8折）
定 价 ¥59.00
评 论 99.2%推荐 353条
配送至 广东广州市海珠区 有货 运费说明 本商品提供礼品包装服务
下周一(4月14日)可送达，请在17小时1分钟内下单并选择 "普通快递送货上门"
丛书名 图灵程序设计丛书
作 者 （奥地利）詹尼士 等著，蒋凡 译
出版社 人民邮电出版社
出版时间 2013-7-1
ISBN 9787115310699
所属分类 图书 > 计算机/网络 > 软件工程/开发项目管理
我要买 1 件
加入购物车
一键购买
收藏商品
收藏人气：1
最佳拍档
¥47.20
推荐系统(推荐系统必读经典，百度技
¥39.20
推荐系统实践(《浪潮之巅》、《数学
¥48.70
机器学习实战【利用Python透析主流
¥45.60
社交网站的数据挖掘与分析（2011年J
¥739.00
【当当金牌店】WD 西部数据 My Passp
¥89.00
【当当自营】logitech罗技 M215二代
新版调查
回顶部

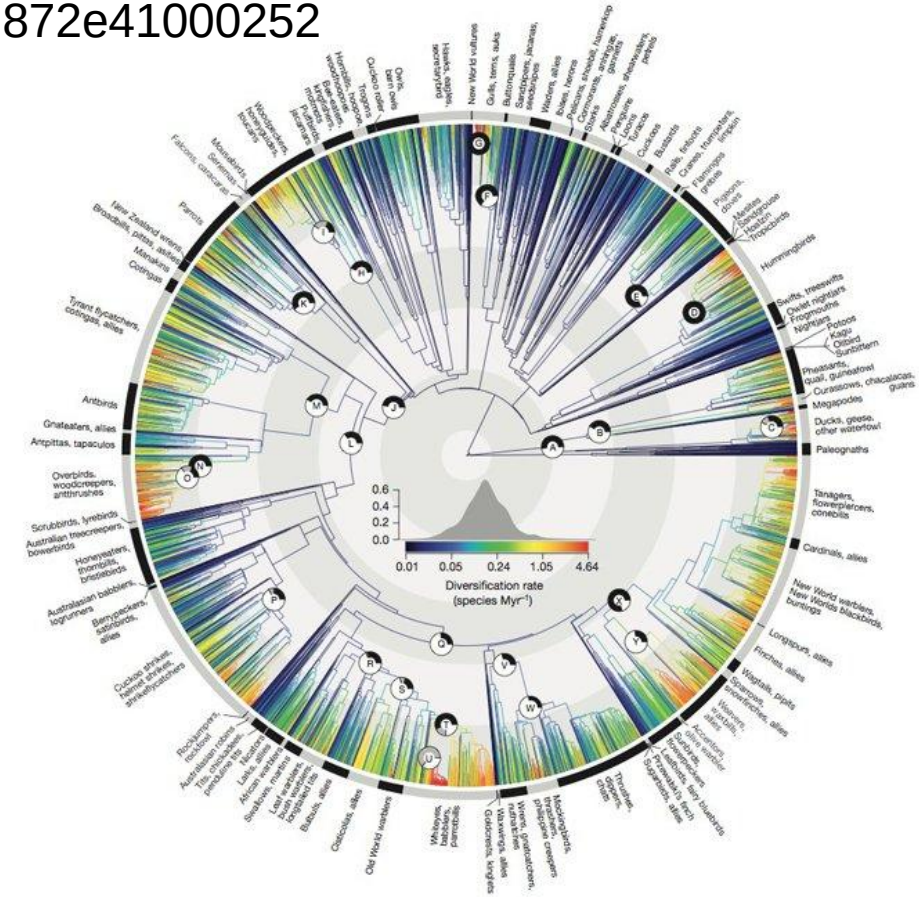
聚类的应用场景：社区发现



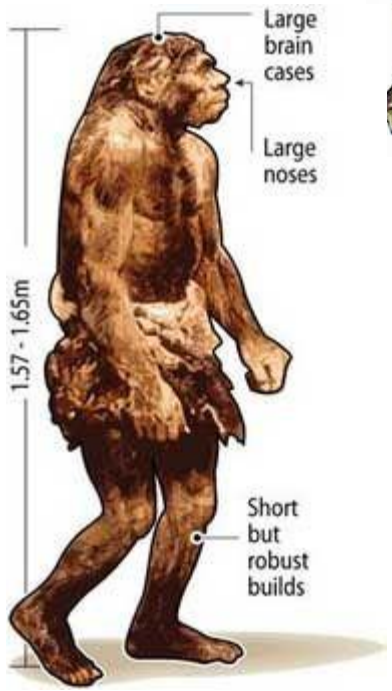
聚类应用场景：生物进化树

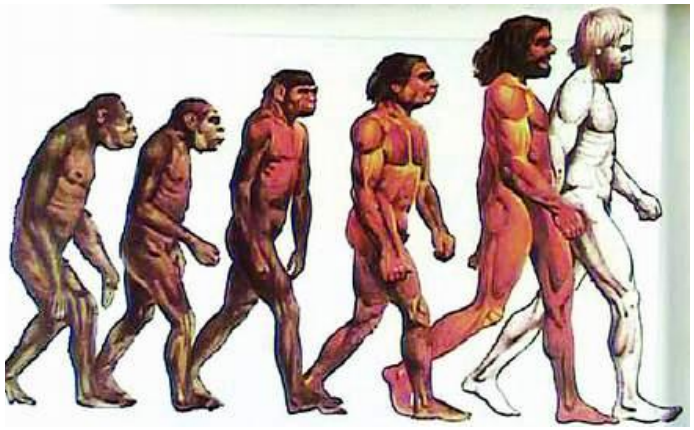


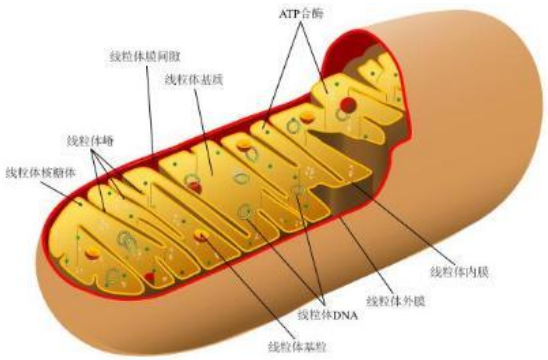
2012最佳科技图表：
http://shijue.me/show_idea/50d129c48ddf872e41000252

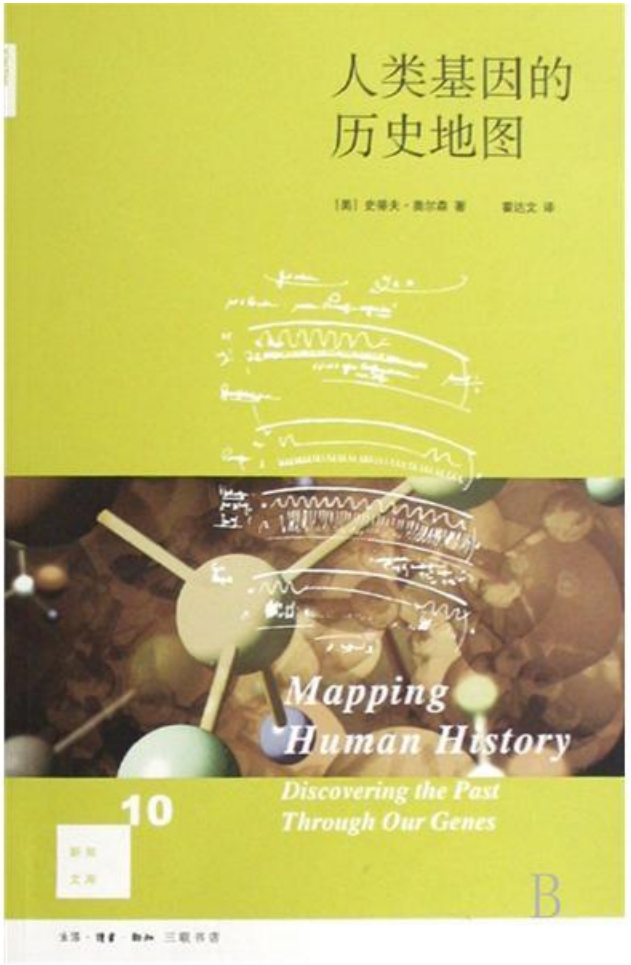


人类基因的历史地图


Large brain cases
Large noses
1.57 - 1.65m
Short but robust builds




ATP合酶
线粒体膜间隙
线粒体基质
线粒体嵴
线粒体核糖体
线粒体内膜
线粒体外膜
线粒体DNA
线粒体基粒


人类基因的
历史地图
Mapping
Human History
Discovering the Past
Through Our Genes
10

聚类应用场景：孤立点的特殊意义

- 信用卡诈骗
- 黑客攻击

```
xmenu=1&inajax=1" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.045
06.30)"
183.3.51.76 - - [29/Nov/2013:01:27:25 +0800] "GET /member.php?mod=logging&action=login HTTP/1.1" 200 17707 "http://r.dataguru
.cn/member.php?mod=logging&action=login" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.
50727; .NET CLR 3.0.04506.30)"
183.3.51.76 - - [29/Nov/2013:01:27:26 +0800] "GET /member.php?mod=logging&action=login HTTP/1.1" 200 17707 "http://r.dataguru
.cn/member.php?mod=logging&action=login" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.
50727; .NET CLR 3.0.04506.30)"
183.3.51.76 - - [29/Nov/2013:01:27:26 +0800] "POST /member.php?mod=logging&action=login&loginsubmit=yes&inajax=1&ajaxmenu=1 H
TTP/1.1" 200 297 "http://r.dataguru.cn/member.php?mod=logging&action=login&loginsubmit=yes&inajax=1&ajaxmenu=1" "Mozilla/4.0
(compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.04506.30)"
66.249.64.1 - - [29/Nov/2013:01:30:19 +0800] "GET /home.php?mod=space&uid=50144&do=home&view=me&from=space HTTP/1.1" 200 5769
 "-" "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X) AppleWebKit/536.26 (KHTML, like Gecko) Version/6.0 Mobile/10A5376
e Safari/8536.25 (compatible; Googlebot-Mobile/2.1; +http://www.google.com/bot.html)"
66.249.64.8 - - [29/Nov/2013:01:30:44 +0800] "GET /space-uid-73446.html HTTP/1.1" 200 4782 "-" "Mozilla/5.0 (compatible; Goog
lebot/2.1; +http://www.google.com/bot.html)"
210.51.177.136 - - [29/Nov/2013:01:35:28 +0800] "GET / HTTP/1.0" 200 46531 "-" "User-Agent: Mozilla/5.0 (compatible; MSIE 6.0
;Windows XP)"
66.249.64.1 - - [29/Nov/2013:01:36:52 +0800] "GET /space-uid-73384.html HTTP/1.1" 200 4776 "-" "Mozilla/5.0 (compatible; Goog
lebot/2.1; +http://www.google.com/bot.html)"
66.249.64.1 - - [29/Nov/2013:01:38:25 +0800] "GET /space-uid-73345.html HTTP/1.1" 200 4434 "-" "Mozilla/5.0 (compatible; Goog
lebot/2.1; +http://www.google.com/bot.html)"
183.3.20.129 - - [29/Nov/2013:01:38:45 +0800] "GET /member.php?mod=logging&action=login HTTP/1.1" 200 17707 "http://r.dataguru
.cn/member.php?mod=logging&action=login" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0
.50727; .NET CLR 3.0.04506.30)"
183.3.20.129 - - [29/Nov/2013:01:38:49 +0800] "GET /member.php?mod=logging&action=login HTTP/1.1" 200 17707 "http://r.dataguru
.cn/member.php?mod=logging&action=login" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0
.50727; .NET CLR 3.0.04506.30)"
183.3.20.129 - - [29/Nov/2013:01:38:49 +0800] "POST /member.php?mod=logging&action=login&loginsubmit=yes&inajax=1&ajaxmenu=1
HTTP/1.1" 200 297 "http://r.dataguru.cn/member.php?mod=logging&action=login&loginsubmit=yes&inajax=1&ajaxmenu=1" "Mozilla/4.0
 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.04506.30)"
[root@class2room web_logs]#
```

关键度量指标：距离

- 距离的定义
- 常用距离（薛毅书P469）

绝对值距离

欧氏距离

闵可夫斯基距离

切比雪夫距离

马氏距离

Lance和Williams距离

离散变量的距离计算

dist( )函数

x1=c(1,2,3,4,5)

x2=c(3,2,1,4,6)

x3=c(5,3,5,6,2)

x=data.frame(x1,x2,x3)

```
> dist(x,method="euclidean")
         1        2        3        4
2 2.449490
3 2.828427 2.449490
4 3.316625 4.123106 3.316625
5 5.830952 5.099020 6.164414 4.582576

> dist(x,method="minkowski")
         1        2        3        4
2 2.449490
3 2.828427 2.449490
4 3.316625 4.123106 3.316625
5 5.830952 5.099020 6.164414 4.582576
> dist(x,method="minkowski",p=5)
         1        2        3        4
2 2.024397
3 2.297397 2.024397
4 3.004922 3.143603 3.004922
5 4.323101 4.174686 5.085057 4.025455
```

dist( )函数

```
> y1=c("F","F","M","F","M")
> y2=c("A","B","B","C","A")
> y3=c(2,3,1,2,3)
> y=data.frame(y1,y2,y3)
> dist(y,method="binary")
  1 2 3 4
2 0
3 0 0
4 0 0 0
5 0 0 0 0
警告信息:
In dist(y, method = "binary") : 强制改变过程中产生了NA
> y1=c(1,0,1,1,0,0,1)
> y2=c(1,0,0,0,1,1,1)
> y3=c(1,1,1,0,0,1,1)
> y=data.frame(y1,y2,y3)
> dist(y,method="binary")
          1         2         3         4         5         6
2 0.6666667
3 0.3333333 0.5000000
4 0.6666667 1.0000000 0.5000000
5 0.6666667 1.0000000 1.0000000 1.0000000
6 0.3333333 0.5000000 0.6666667 1.0000000 0.5000000
7 0.0000000 0.6666667 0.3333333 0.6666667 0.6666667 0.3333333
```

数据中心化与标准化变换

- 目的：使到各个变量平等地发挥作用
- scale( )函数
- 极差化。 sweep( )函数（薛毅书P473）

```
> x
  x1 x2 x3
1  1  3  5
2  2  2  3
3  3  1  5
4  4  4  6
5  5  6  2
> scale(x,center=TRUE,scale=TRUE)
           x1         x2         x3
[1,] -1.2649111 -0.1039750  0.4868645
[2,] -0.6324555 -0.6238503 -0.7302967
[3,]  0.0000000 -1.1437255  0.4868645
[4,]  0.6324555  0.4159002  1.0954451
[5,]  1.2649111  1.4556507 -1.3388774
attr(,"scaled:center")
 x1  x2  x3
3.0 3.2 4.2
attr(,"scaled:scale")
      x1       x2       x3
1.581139 1.923538 1.643168
```

对变量进行分类的指标：相似系数

- 距离：对样本进行分类
- 相似系数：对变量进行分类
- 常用相似系数：夹角余弦，相关系数（薛毅书P475）

（凝聚的）层次聚类法

- 思想

1 开始时，每个样本各自作为一类

2 规定某种度量作为样本之间的距离及类与类之间的距离，并计算之

3 将距离最短的两个类合并为一个新类

4 重复2-3，即不断合并最近的两个类，每次减少一个类，直至所有样本被合并为一类

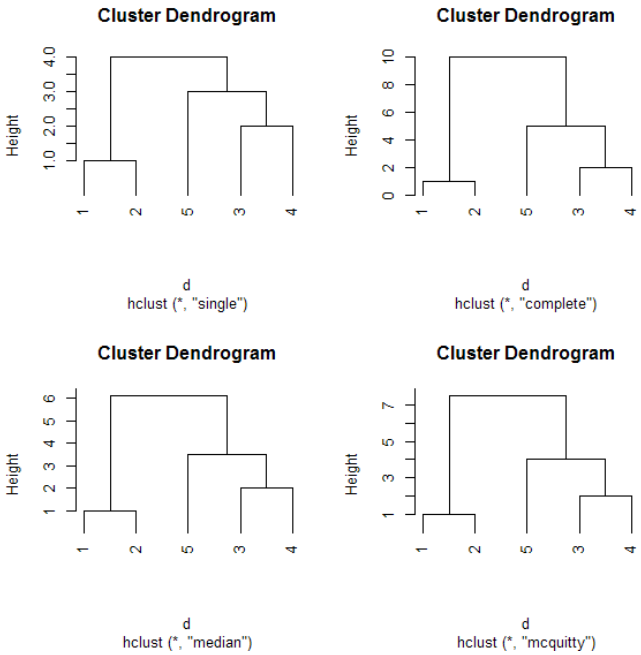
各种类与类之间距离计算的方法

- 薛毅书P476
- 最短距离法
- 最长距离法
- 中间距离法
- 类平均法
- 重心法
- 离差平方和法

hclust( )函数

- 简单的例子（薛毅书P480）

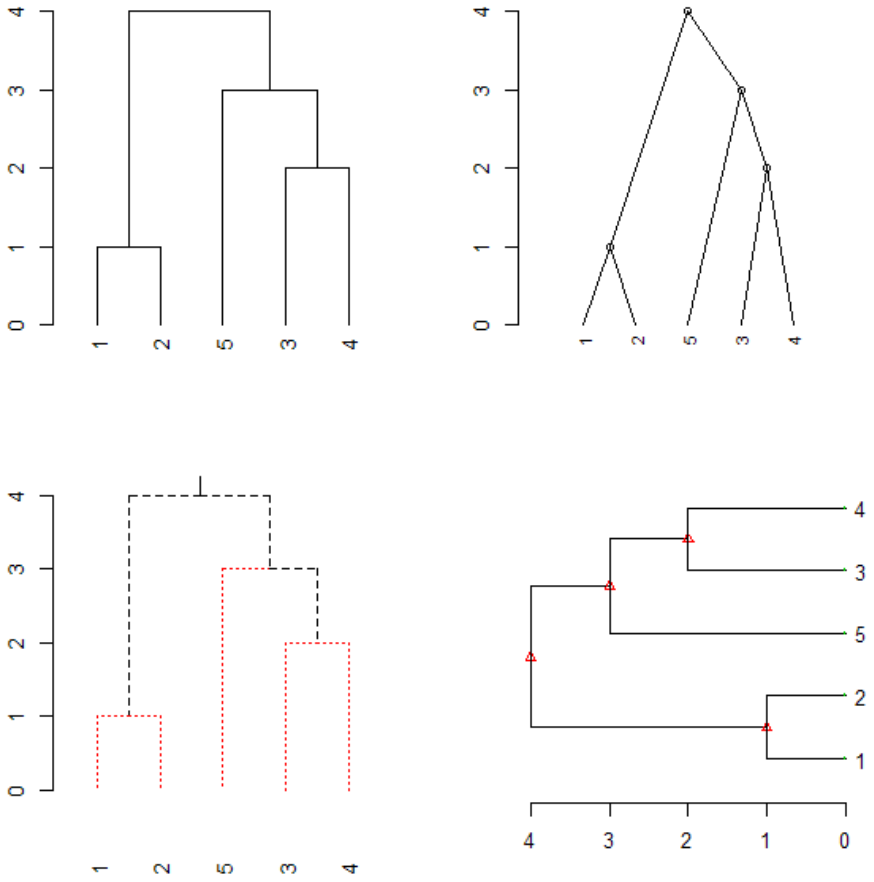
```
> x<-c(1,2,6,8,11); dim(x)<-c(5,1);
> x
     [,1]
[1,]    1
[2,]    2
[3,]    6
[4,]    8
[5,]   11
> d<-dist(x)
> d
   1  2  3  4
2  1
3  5  4
4  7  6  2
5 10  9  5  3
> hc1<-hclust(d, "single"); hc2<-hclust(d, "complete")
> hc3<-hclust(d, "median"); hc4<-hclust(d, "mcquitty")
> opar <- par(mfrow = c(2, 2))
> plot(hc1,hang=-1); plot(hc2,hang=-1)
> plot(hc3,hang=-1); plot(hc4,hang=-1)
> par(opar)
```



各种谱系图画法

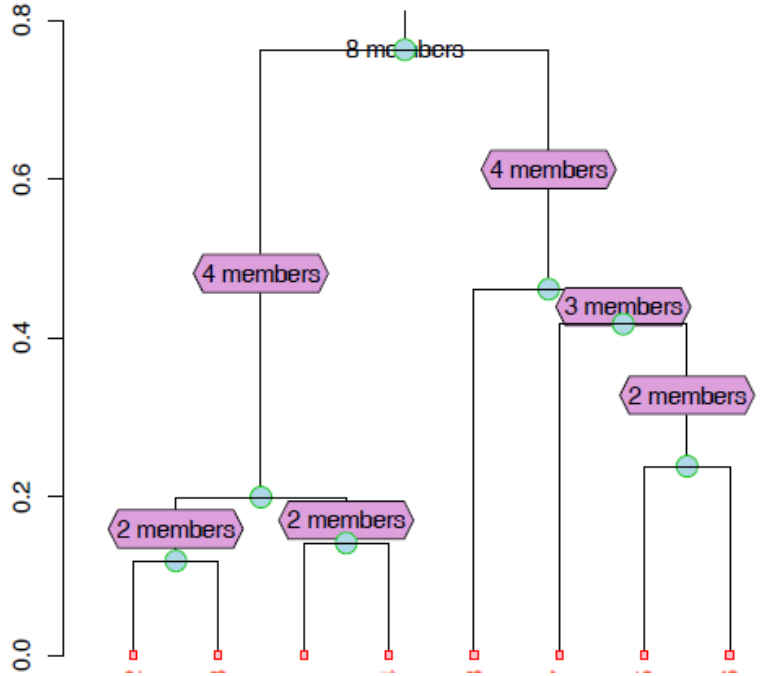
- as.dendrogram( )函数（薛毅书P482）

```
dend1<-as.dendrogram(hc1)
opar <- par(mfrow = c(2, 2),mar = c(4,3,1,2))
plot(dend1)
plot(dend1, nodePar=list(pch = c(1,NA),
     cex=0.8, lab.cex=0.8),
type = "t", center=TRUE)
plot(dend1, edgePar=list(col = 1:2, lty = 2:3),
dLeaf=1, edge.root = TRUE)
plot(dend1, nodePar=list(pch = 2:1,
     cex=.4*2:1, col=2:3),
horiz=TRUE)
par(opar)
```



对变量进行聚类分析

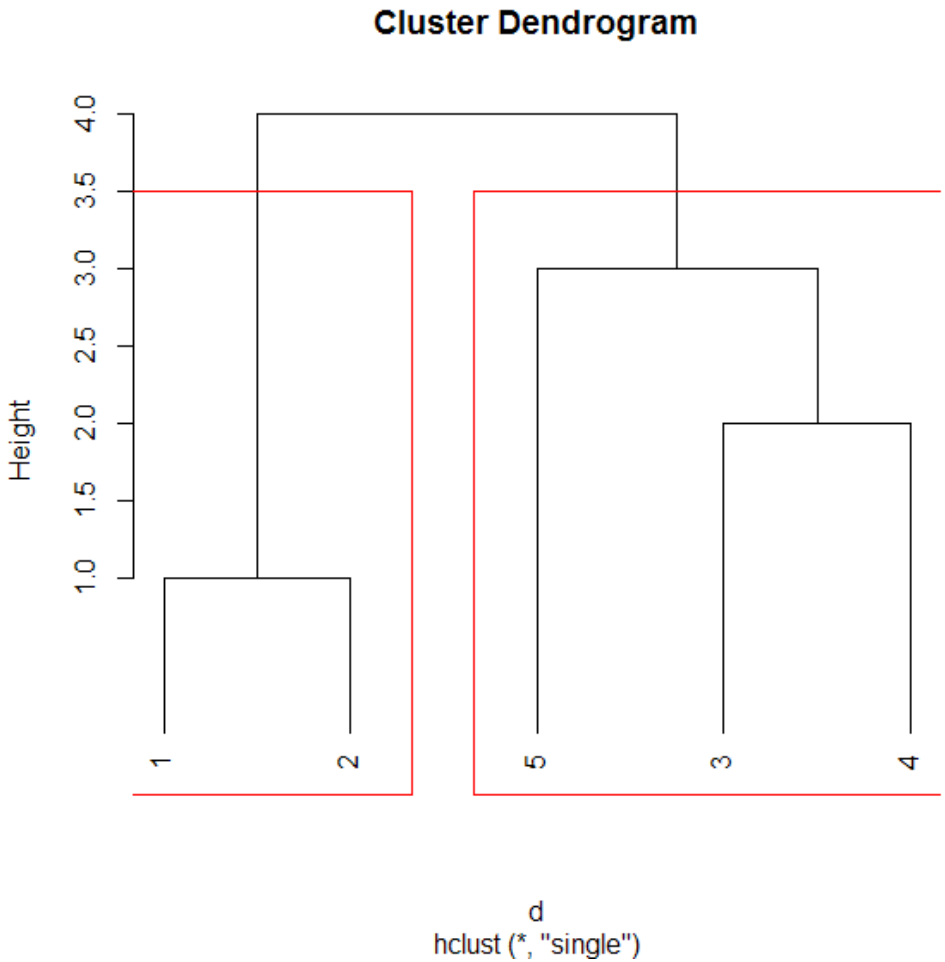
- 例子（薛毅书P483）



分多少个类？

- rect.hclust( )函数

```
> plot(hc1,hang=-1)
> rect.hclust(hc1,k=2)
```



综合实例

- 薛毅书P487

动态聚类：K-means方法

- 算法：

1 选择K个点作为初始质心

2 将每个点指派到最近的质心，形成K个簇（聚类）

3 重新计算每个簇的质心

4 重复2-3直至质心不发生变化

kmeans( )函数

```
> X=iris[,1:4]
> km=kmeans(X,3)
>
>
> km
K-means clustering with 3 clusters of sizes 62, 50, 38

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.901613    2.748387     4.393548    1.433871
2     5.006000    3.428000     1.462000    0.246000
3     6.850000    3.073684     5.742105    2.071053

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [37] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [73] 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 3 3 3 3 1 3
[109] 3 3 3 3 3 1 1 3 3 3 3 1 3 1 3 1 3 3 1 1 3 3 3 3 3 1 3 3 3 3 1 3 3 3
[145] 3 3 1 3 3 1
```

K-means算法的优缺点

- 有效率，而且不容易受初始值选择的影响
- 不能处理非球形的簇
- 不能处理不同尺寸，不同密度的簇
- 离群值可能有较大干扰（因此要先剔除）

基于有代表性的点的技术：K中心聚类法

- 算法步骤

1 随机选择k个点作为“中心点”

2 计算剩余的点到这k个中心点的距离，每个点被分配到最近的中心点组成聚簇

3 随机选择一个非中心点O_r，用它代替某个现有的中心点O_j，计算这个代换的**总代价S**

4 如果S<0，则用O_r代替O_j，形成新的k个中心点集合

5 重复2，直至中心点集合不发生变化

K中心法的实现：PAM

- PAM使用离差平方和来计算成本S（类似于ward距离的计算）
- R语言的cluster包实现了PAM
- K中心法的优点：对于“噪音较大和存在离群值的情况，K中心法更加健壮，不像Kmeans那样容易受到极端数据影响
- K中心法的缺点：执行代价更高

cluster包的pam( )函数

```
> x=iris[,1:4]
> kc=pam(x,3)
> kc
Medoids:
      ID Sepal.Length Sepal.Width Petal.Length Petal.Width
[1,]   8          5.0         3.4          1.5         0.2
[2,]  79          6.0         2.9          4.5         1.5
[3,] 113          6.8         3.0          5.5         2.1
Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [37] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [73] 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3
[109] 3 3 3 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3 3 3 2 3
[145] 3 3 2 3 3 2
Objective function:
    build      swap
0.6709391 0.6542077

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
 [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

CLARA

- Cluster LARge Apllication，用于对大数据集进行快速聚类
- 大数据处理的三种基本思路，关键字：抽样，精度，性能
- 算法思想：

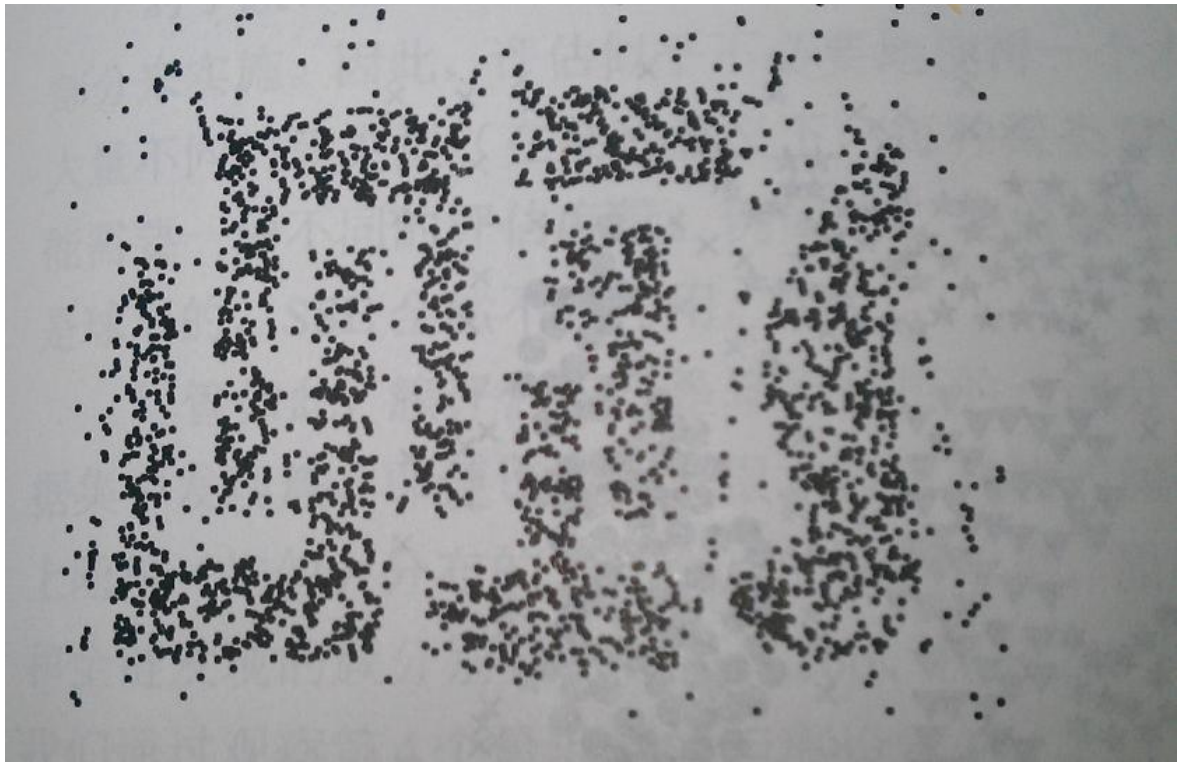
1 从大数据集中抽取少量样本

2 对抽取出来的样本进行PAM聚类

3 从步骤2可以获得聚类中心，使用这组聚类中心对大数据集进行聚类，分类原则是按样本点离各聚类中心距离最短者划分簇

基于密度的方法: DBSCAN

- DBSCAN = Density-Based Spatial Clustering of Applications with Noise
- 本算法将具有**足够高密度**的区域划分为簇，并可以发现**任何形状**的聚类



若干概念

r-邻域：给定点半径**r**内的区域

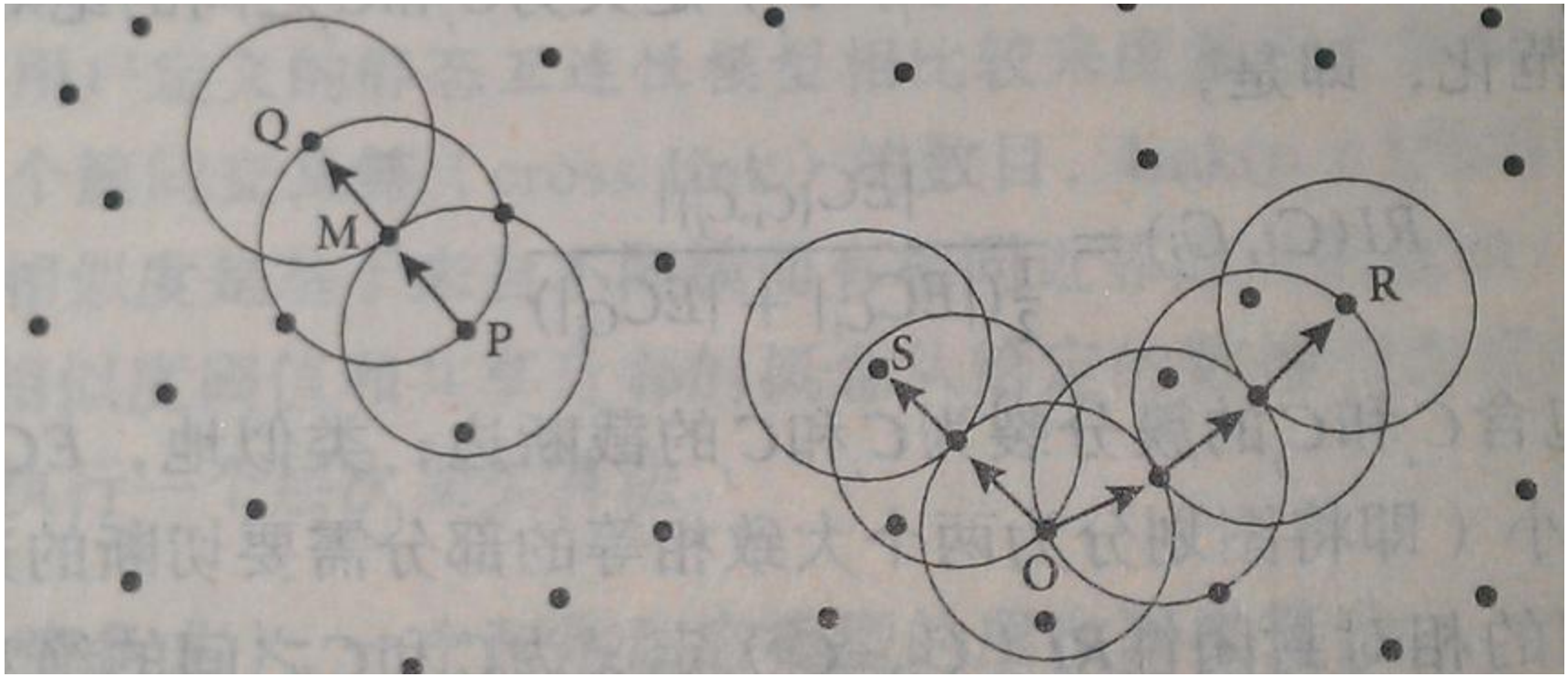
核心点：如果一个点的r-邻域至少包含最少数目**M**个点，则称该点为核心点

直接密度可达：如果点p在核心点q的r-邻域内，则称p是从q出发可以直接密度可达

如果存在点链$p_1,p_2, ..., p_n$，$p_1=q$，$p_n=p$，p_{i+1}是从p_i关于r和M直接密度可达，则称点p是从q关于r和M**密度可达**的

如果样本集D中存在点o，使得点p、q是从o关于 r和M密度可达的，那么点p、q是关于r和M**密度相连**的

DBSCAN



DBSCAN

- 算法基本 思想

1 指定合适的 r 和 M

2 计算所有的样本点，如果点p的r邻域里有超过M个点，则创建一个以p为核心点的新簇

3 反复寻找这些核心点直接密度可达（之后可能是密度可达）的点，将其加入到相应的簇，对于核心点发生“密度相连”状况的簇，给予合并

4 当没有新的点可以被添加到任何簇时，算法结束

DBSCAN算法描述

输入: 包含n个对象的数据库，半径e，最少数目MinPts;

输出:所有生成的簇，达到密度要求。

(1)Repeat

(2)从数据库中抽出一个未处理的点；

(3)IF抽出的点是核心点 THEN 找出所有从该点密度可达的对象，形成一个簇；

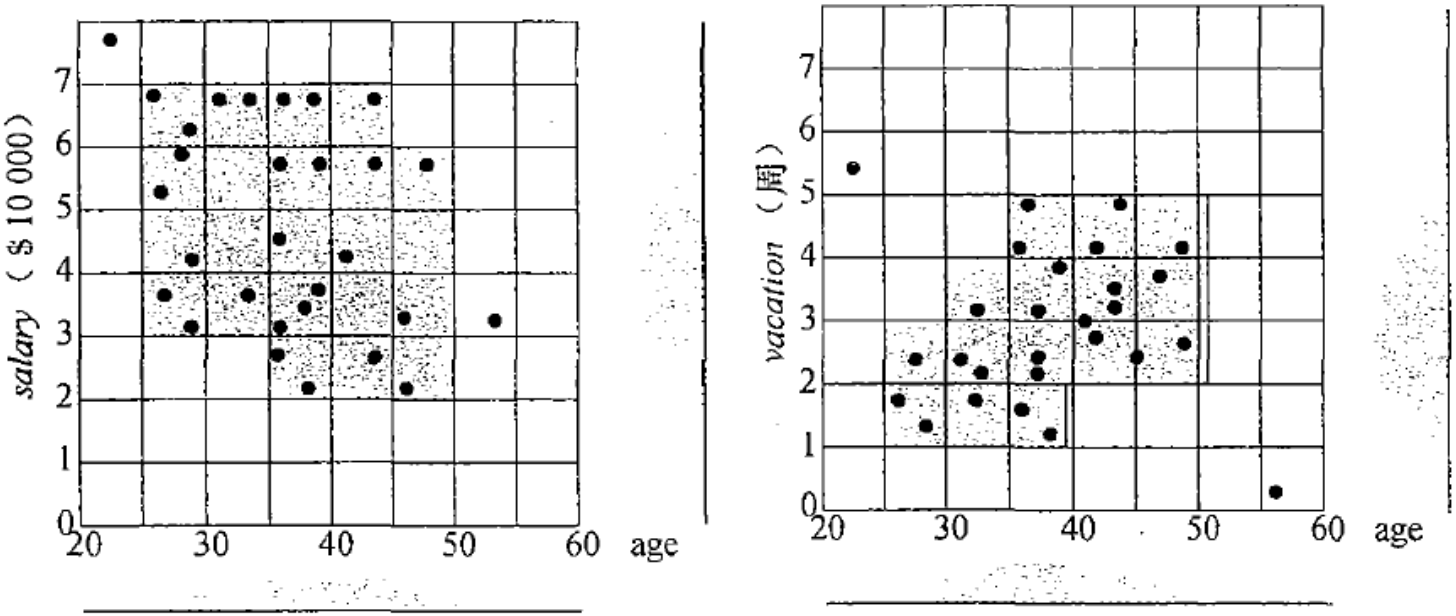
(4)ELSE 抽出的点是边缘点(非核心对象)，跳出本次循环，寻找下一个点；

(5)UNTIL 所有的点都被处理。

DBSCAN对用户定义的参数很敏感，细微的不同都可能导致差别很大的结果，而参数的选择无规律可循，只能靠经验确定。

基于网格的聚类方法CLIQUE

- Clustering In QUEst，clique本身的词义是“小圈子，门派，阀”
- 基于网格的聚类方法，可以发现基于密度的簇
- 需要给出两个参数：一个是网格的步长，一具是密度阈值
- 使用类似关联规则挖掘中的Apriori算法的先验性质（见下图）



稠密单元

- 什么是稠密单元？
- 通过对低维度识别稠密单元，获得高维度下的候选稠密单元
- 只对候选稠密单元进行筛选，从而降低了计算量

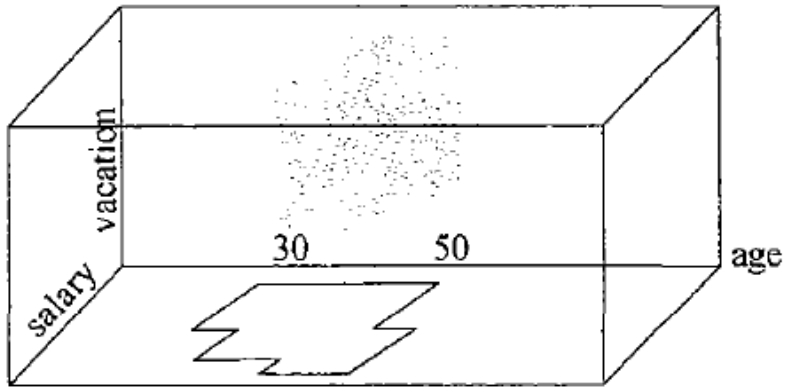
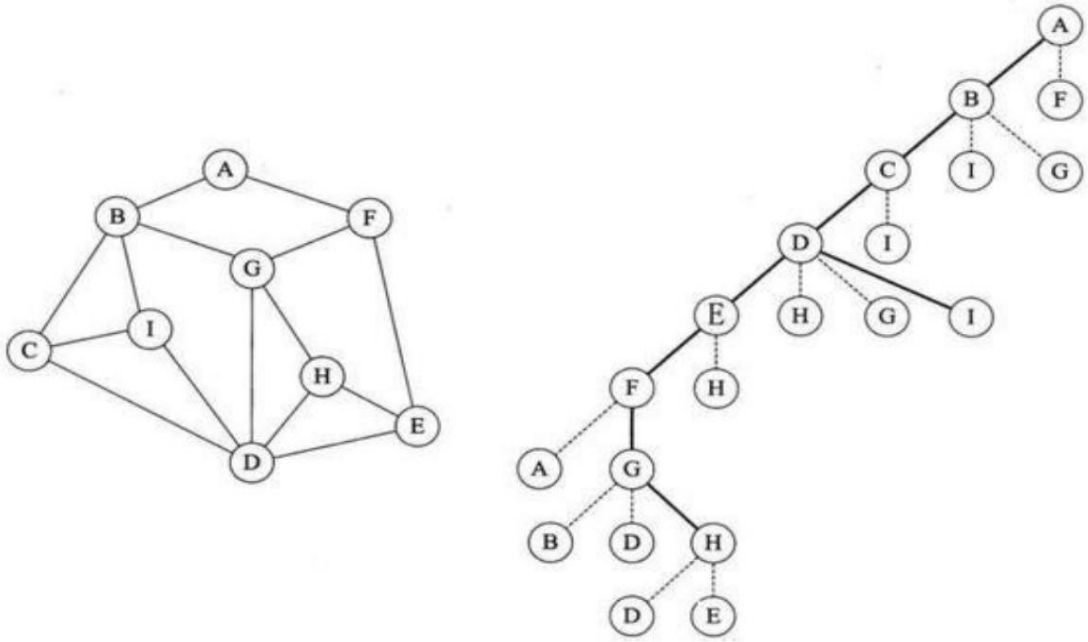


图 10.20　对 *salary* 和 *vacation* 维上发现的关于 *age* 的稠密单元取交，从而为发现更高维度上的稠密单元提供候选搜索空间

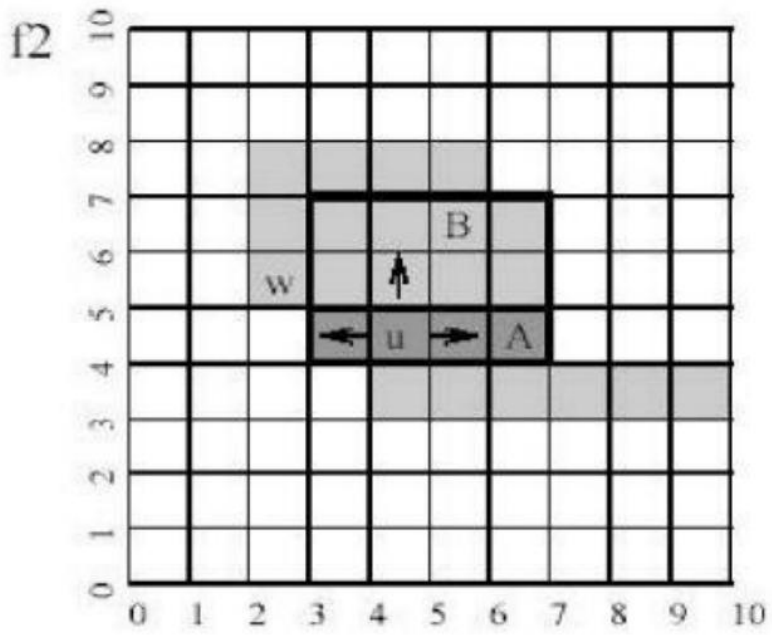
簇的装配过程

- 邻接网格：满足四方连续或八方连续的网格称为邻接网格。实际应用中常常根据实际数据的数量和密集程度等条件选择不同的网格连续性
- 深度优先遍历：顶点——稠密单元，边——邻接关系
- 遍历完成后，簇被完全装配

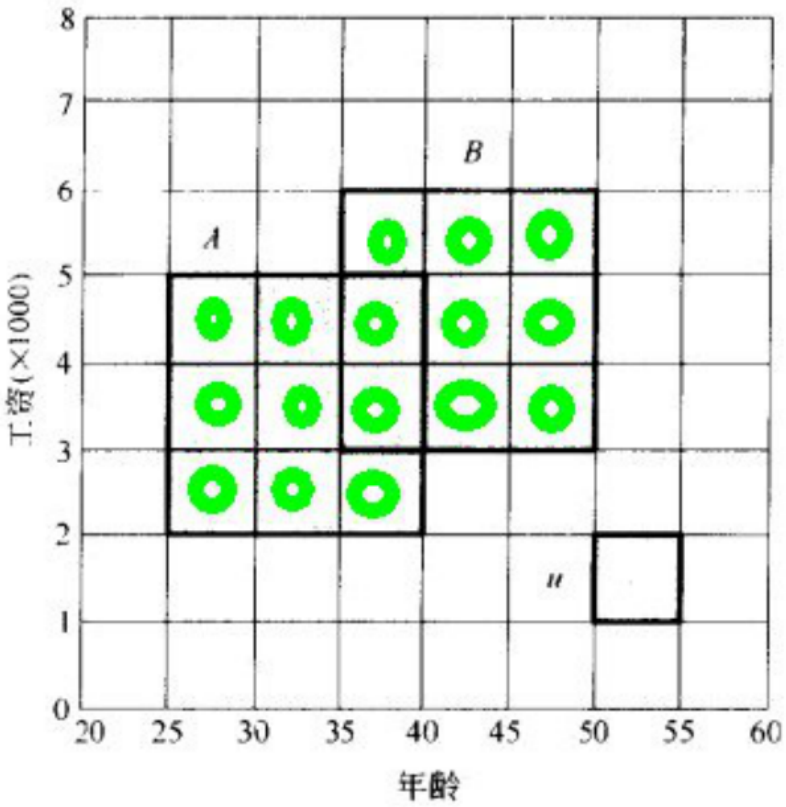


生成簇的最小化描述

- 把稠密单元合并为“最大区域”
- 最大区域是超矩形，落入该区域的单元都是稠密的，并且在各个维度上都无法再扩展
- 贪心算法



最小覆盖



图为类的最小描述示意图。由年龄和工资构成的两维空间分割为8×8的网格，每一个网格代表一个单元，如u= 50≤年龄≤55）∧（1≤工资≤2）。其中A和B两个区域分别为A=（25≤年龄≤40）∧（2≤工资≤5），B=（35≤年龄≤50）∧（3≤工资≤6）。假设高密集区域用阴影表示，则A∪B形成了一个类。那么该类的最小表示为（（25≤年龄≤40）∧（2≤工资≤5））∨（（35≤年龄≤50）∧（3≤工资≤6））。这里A和B分别表示两个最大区域覆盖，最后得到的类的表示即为最小覆盖。 实际中则将2维拓展到k维即可。

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间