数据挖掘技术
聚类分析:基本概念
和算法
第七章的课堂笔记
数据挖掘导论，第二版
经过
谭、斯坦贝克、卡帕特内、库马尔

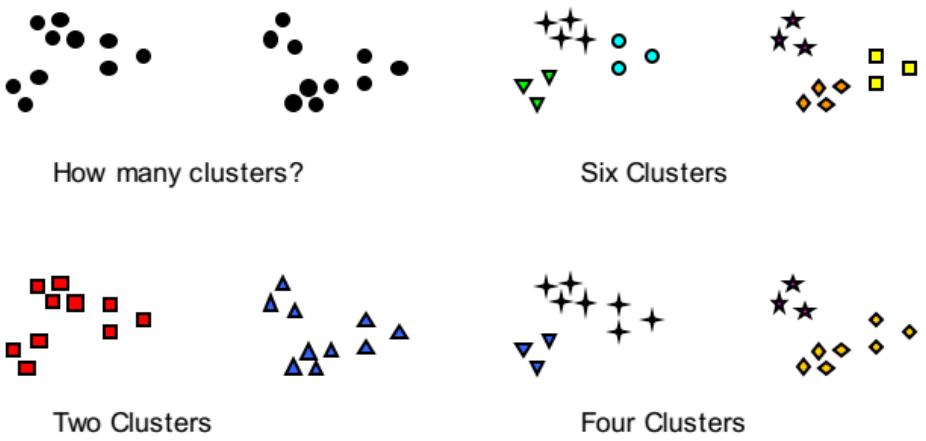
02/14/2018 数据挖掘导论，第 2 版 2
什么是聚类分析?
●寻找对象组，使得组中的对象
彼此相似(或相关),与其他组中的对象不同(或无关)
集群间
距离是
最大化
集群内
距离是
最小化

02/14/2018 数据挖掘导论，第 2 版 3
聚类分析的应用
●理解
组相关文档
为了浏览，将基因分组
和蛋白质
相似的功能，或具有相似价格波动的组合股票
●总结
减小大的尺寸
数据集
发现集群产业群
1 应用向下传输、向下传输、向下传输、向下传输、向下传输、向下传输、向下传输、向下传输、向下传输、向下传输、向下传输。
MicronTechDOWN，TexasInstDown，TellabsIncDown，NatlSemiconductDOWN，OraclDOWN，SGIDOWN，Sun-DOWN
技术 1 拥有
2 个苹果电脑关机、自动桌面关
机、DECDOWN、ADVMicroDeviceDOWN、AndrewCorpDOWN、计算机关联关机、电路城市关机、
CompaqDOWN、EMCCorpDOWN、GenInstDOWN、MotorolaDOWN、MicrosoftDOWN、ScientificAtlDOWN
技术 2 拥有
3 FannieMaeDOWN，FedHomeLoanDOWN，MBNACorpDOWN，MorgansTanlyDown Financial-Down 4
BakerHughеSup，DresserIndsUP，HalliburtonHLDUP，LouisianaLandUP，PhillipsPetroUP，UnocalUP，Schlumberger-UP OilUP
聚类降水

在澳大利亚

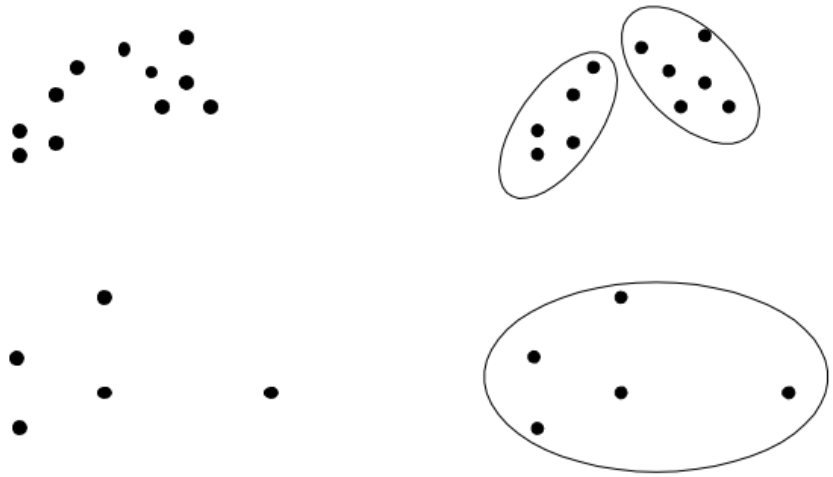
02/14/2018 数据挖掘导论，第 2 版 4
什么不是聚类分析?
●简单分割
将学生分成不同的注册组
按姓氏的字母顺序
●查询结果
分组是外部规范的结果
聚类是基于数据的对象分组
●监督分类
有班级标签信息
●关联分析
本地与全球连接

Notion of a Cluster can be Ambiguous



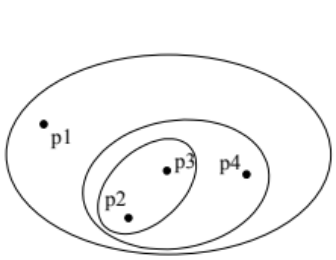
02/14/2018 数据挖掘导论，第 2 版 6
集群的类型
●集群是一组集群
●聚类的层次集和分区集之间的重要区别●分区聚类
将数据对象划分为不重叠的子集
(集群)使得每个数据对象恰好在一个子集中
●分层聚类
组织成层次树的一组嵌套集群

Partitional Clustering

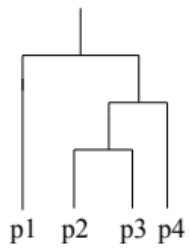


Original Points　　A Partitional Clustering

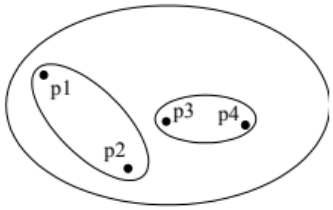
Hierarchical Clustering



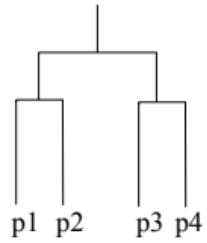
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



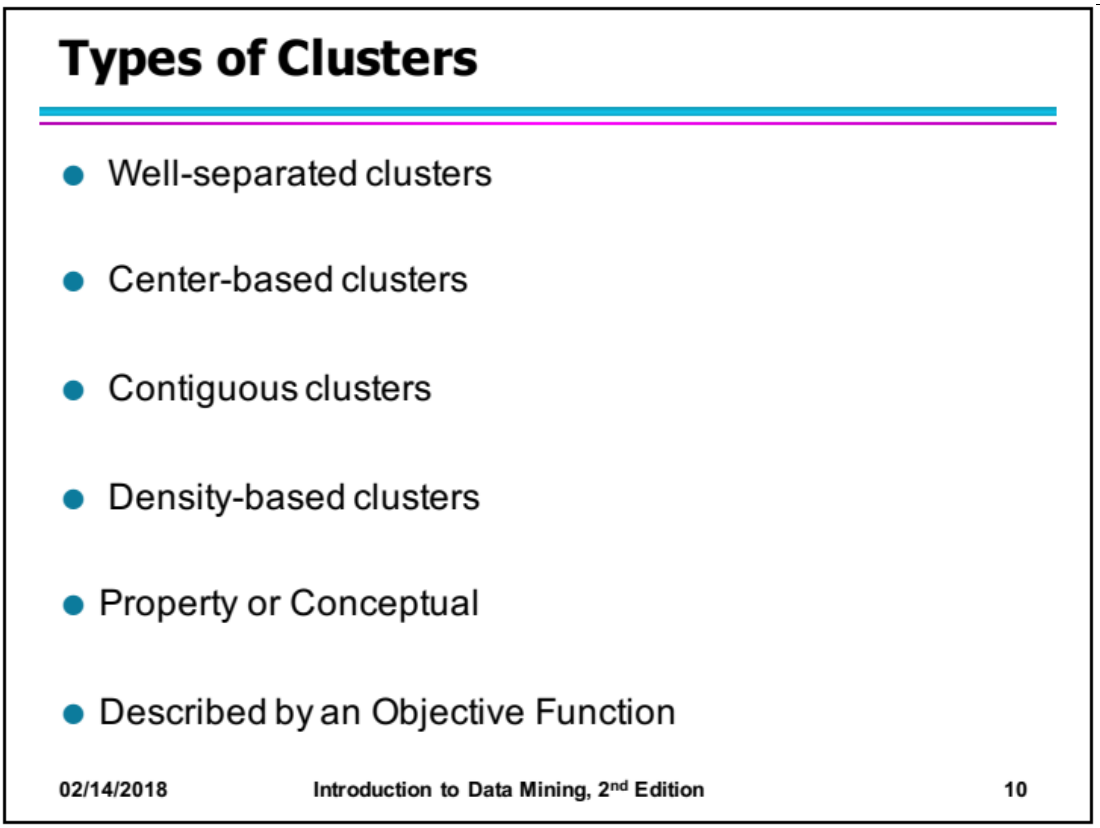
Non-traditional Dendrogram

02/14/2018 数据挖掘导论，第 2 版 9

聚类集之间的其他区别

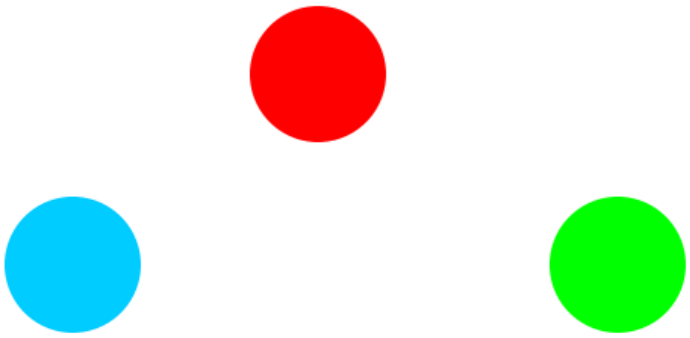
●排他性与非排他性

在非排他性聚类中，点可能属于多个聚类。
可以表示多个类或“边界”点
●模糊与非模糊
在模糊聚类中，一个点属于每个权重在 0 和 1 之间的聚类。权重必须等于 1
概率聚类具有相似的特征
●部分与全部
在某些情况下，我们只想对一些数据进行聚类
●异质与同质
不同大小、形状和密度的集群



Types of Clusters: Well-Separated

- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



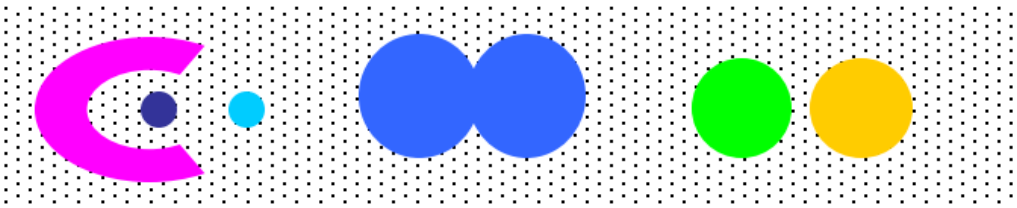
3 well-separated clusters

02/14/2018 数据挖掘导论，第 2 版 12
集群类型:基于中心
●基于中心
集群是一组对象，集群中的对象
更接近(更相似)一个集群的“中心”,而不是任何其他集群的中心
聚类的中心通常是质心，即聚类中所有点的平均值，或者是中间点，即聚类中最具“代表性”的点
4 个基于中心的集群

02/14/2018 数据挖掘导论，第 2 版 13
集群类型:基于邻接
●相邻群集(最近的邻居或
可传递)
聚类是一组点，使得聚类中的一个点
比不在集群中的任何点更接近(或更相似)集群中的一个或多个其他点。
8 个连续的集群

Types of Clusters: Density-Based

- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.

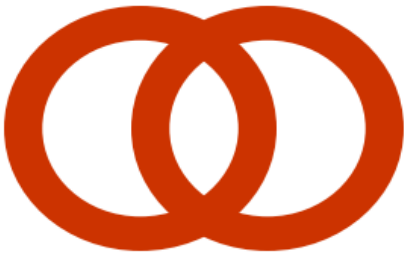


6 density-based clusters

Types of Clusters: Conceptual Clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.

.



2 Overlapping Circles

02/14/2018 数据挖掘导论，第 2 版 16
聚类类型:目标函数
●由目标函数定义的聚类
查找最小化或最大化目标函数的聚类。
列举所有可能的将点分成簇的方法
使用以下方法评估每组潜在聚类的“良好性”
给定的目标函数。(NP 难)可以有全球或本地目标。
分层聚类算法通常有局部目标
全局目标函数方法的一个变体是将数据拟合到参数化模型。
u 模型参数由数据确定。
混合模型假设数据是一系列“混合”的
统计分布。

02/14/2018 数据挖掘导论，第 2 版 17
将聚类问题映射到不同的问题
●将聚类问题映射到不同的领域，并解决该领域中的相关问题
邻近矩阵定义了一个加权图，其中
节点是被聚集的点，而
加权边代表了
点
聚类相当于将图分解成连接的组件，每个组件对应一个聚类。
想要最小化集群之间的边缘权重
并且最大化聚类内的边缘权重

输入数据的特征很重要
●接近或密度测量的类型
集群的核心
取决于数据和应用
●影响接近度和/或密度的数据特征有
维度
稀疏
属性类型
数据中的特殊关系
u 例如，自相关
数据的分发
●噪声和异常值
经常干扰聚类算法的操作

Clustering Algorithms

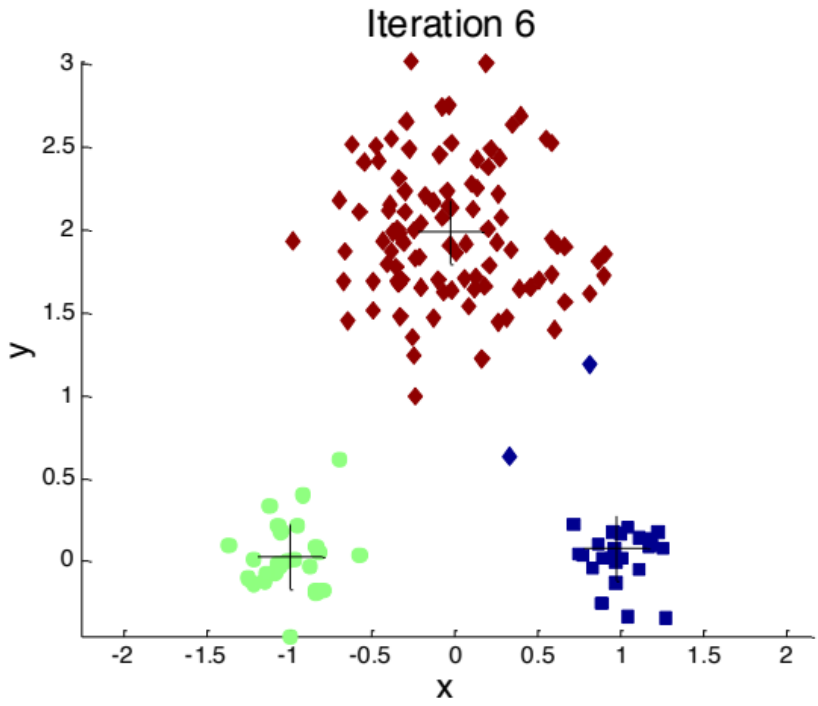
- K-means and its variants
- Hierarchical clustering
- Density-based clustering

K-means Clustering

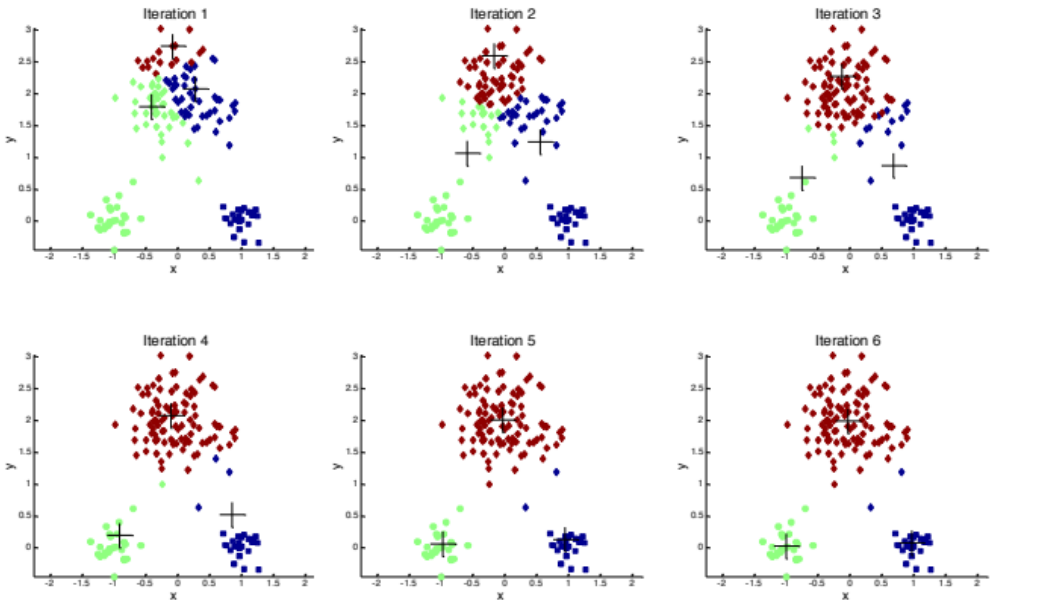
- Partitional clustering approach
- Number of clusters, K, must be specified
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- The basic algorithm is very simple

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Example of K-means Clustering



Example of K-means Clustering



02/14/2018 数据挖掘导论，第 2 版 23

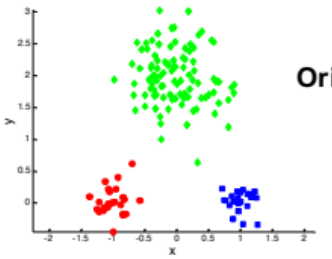
Kmeans 群集详细信息

●初始质心通常是随机选择的。

每次运行产生的集群各不相同。
●质心(通常)是集群中点的平均值。
●“接近度”用欧几里德距离、余弦来衡量
相似性、相关性等。
●对于上述常见的相似性度量，知识分子将会收敛。
●大部分趋同发生在前几个阶段
迭代。
通常停止条件被改变为“直到相对较少的点改变簇”
●复杂度为 0(n * K * I * d)
n =点数，K =聚类数，
I =迭代次数，d =属性数

02/14/2018 数据挖掘导论，第 2 版 24
评估知识管理集群
●最常见的度量是平方误差之和(SSE)
对于每个点，误差是到最近聚类的距离
为了得到 SSE，我们将这些误差平方并求和。
x 是集群 Ci 中的数据点，mi 是
群集 Ci
u 可以显示 mi 对应于聚类的中心(平均值)
给定两组聚类，我们更喜欢最小的一组
错误
减少上交所的一个简单方法是增加 K，即
簇
具有较小 K 值的好聚类比具有较高 K 值的差聚类具有较低的 SSE
∑=∑=
K
i x C
i
i
上交所区 m x
1
(，)

Two different K-means Clusterings

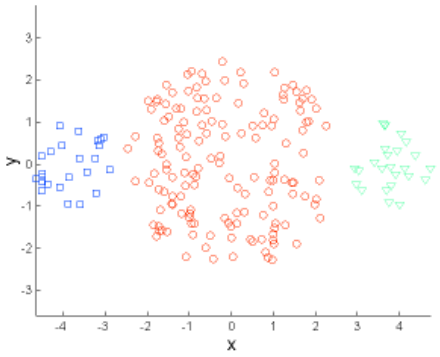


Original Points

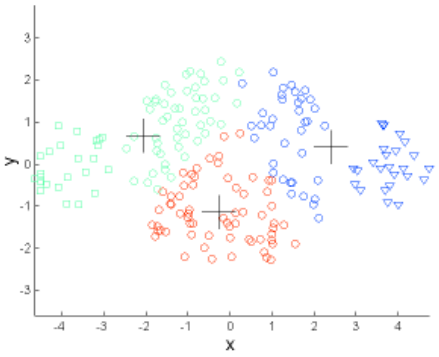
Limitations of K-means

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes
- K-means has problems when the data contains outliers.

Limitations of K-means: Differing Sizes

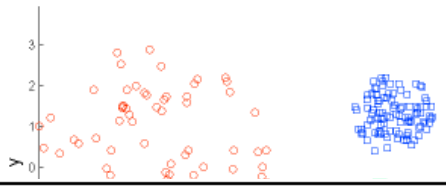


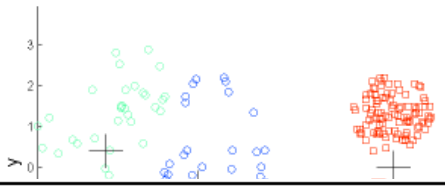
Original Points



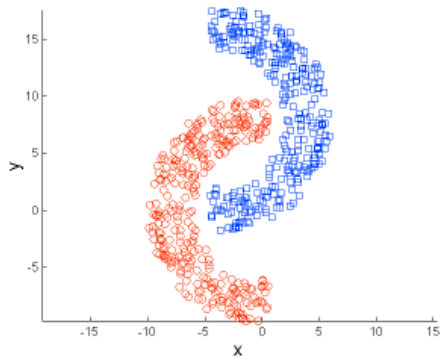
K-means (3 Clusters)

Limitations of K-means: Differing Density

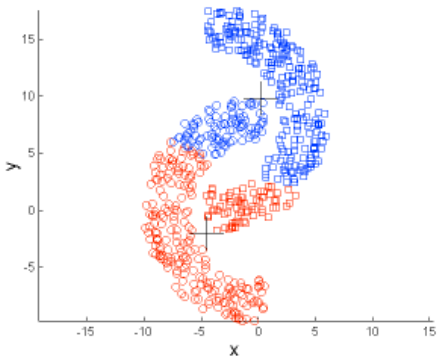




Limitations of K-means: Non-globular Shapes

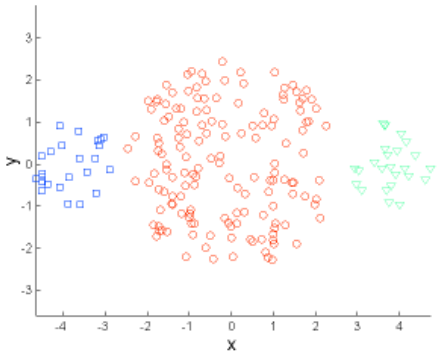


Original Points

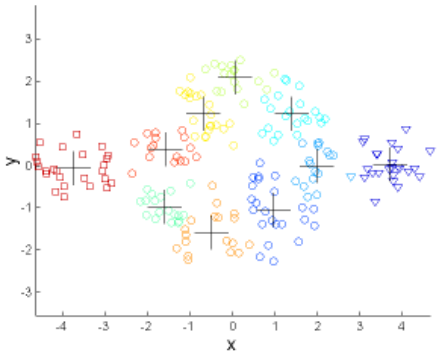


K-means (2 Clusters)

Overcoming K-means Limitations



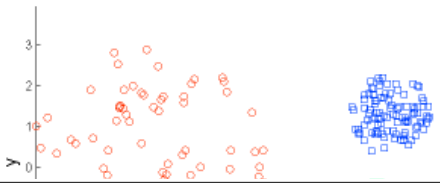
Original Points



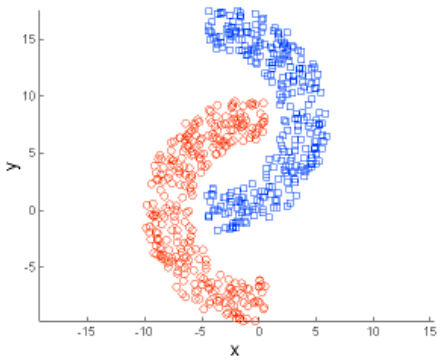
K-means Clusters

One solution is to use many clusters.
Find parts of clusters, but need to put together.

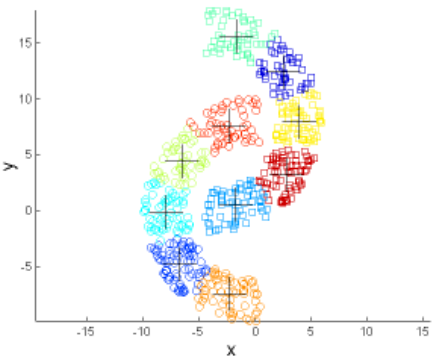
Overcoming K-means Limitations



Overcoming K-means Limitations

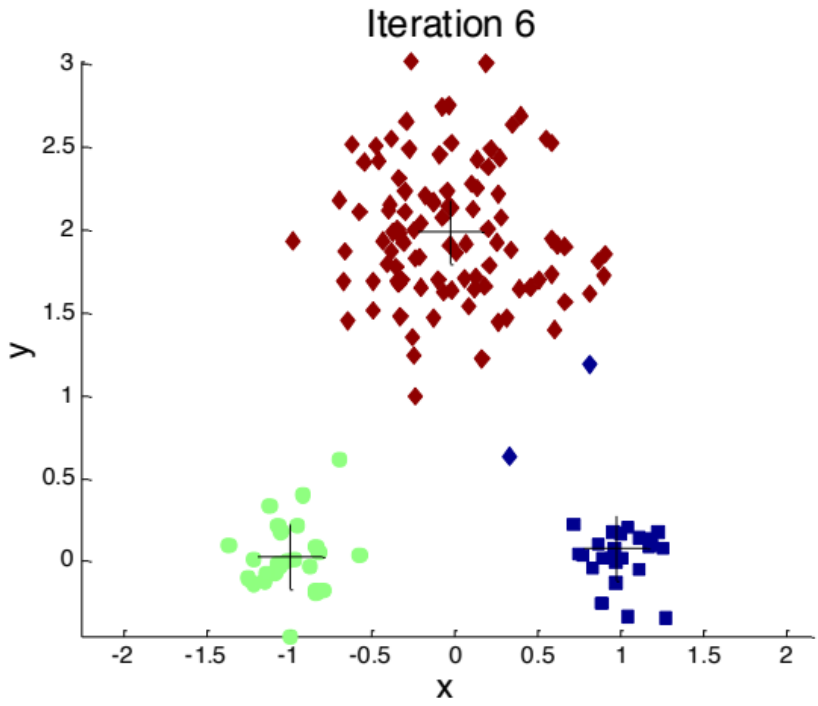


Original Points

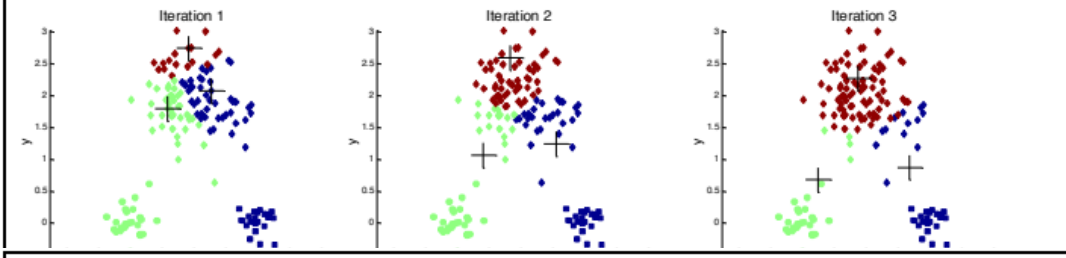


K-means Clusters

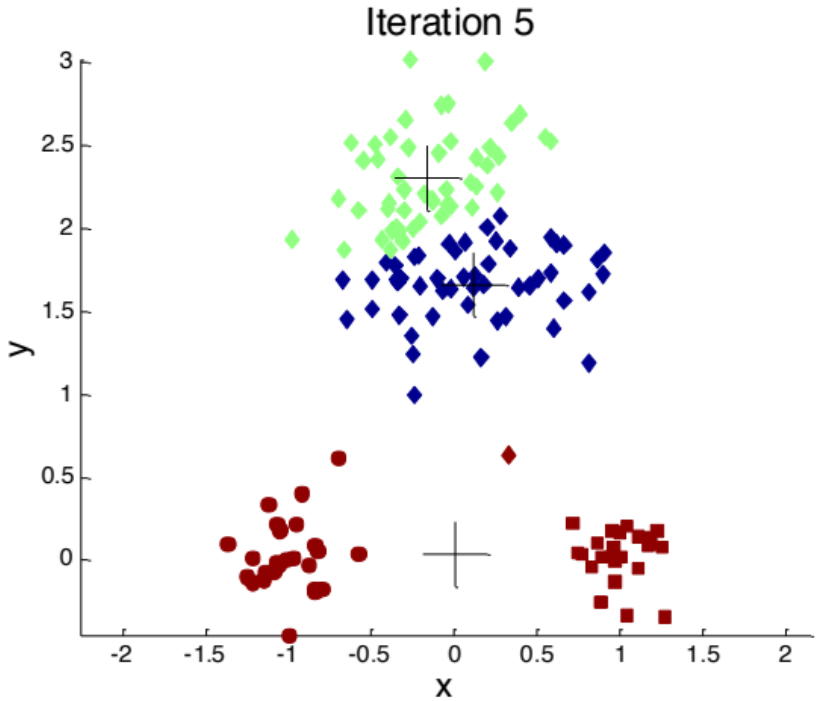
Importance of Choosing Initial Centroids



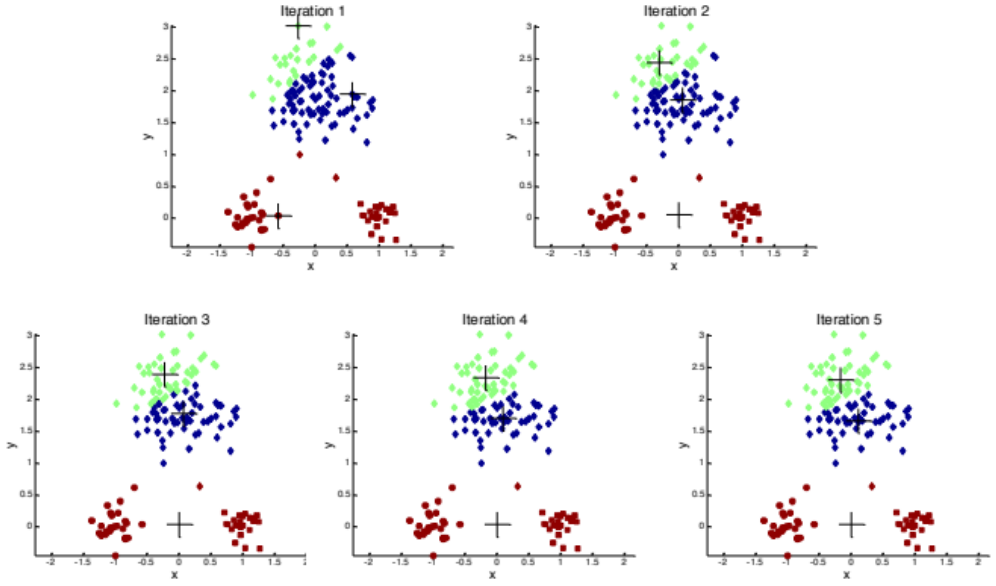
Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids ...



Importance of Choosing Initial Centroids …



02/14/2018 数据挖掘导论，第 2 版 37

选择初始点的问题

●如果有 K 个“真实”聚类，那么从每个聚类中选择一个质心的机会很小。

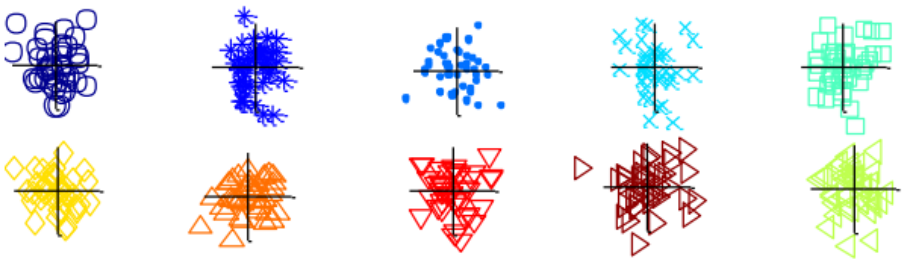
当 K 较大时，机会相对较小。如果簇大小相同，则

例如，如果 K = 10，那么概率= 10！/10= 0.00036 有时初始质心会重新调整

正确的方式，有时他们不会

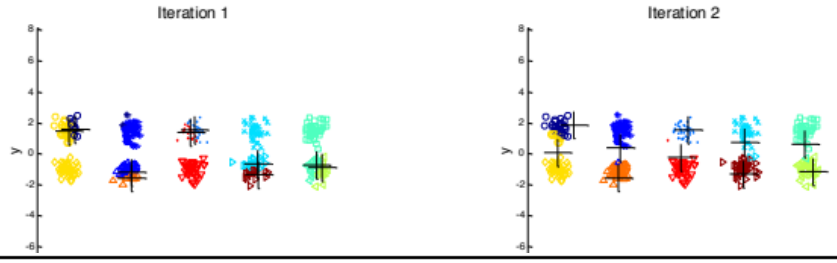
考虑一个五对集群的例子

10 Clusters Example

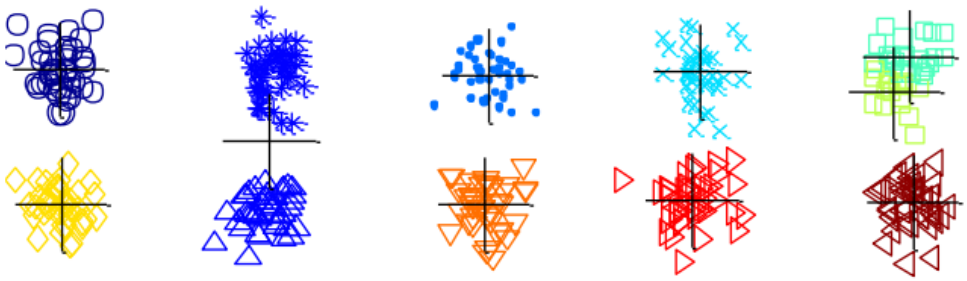


Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example

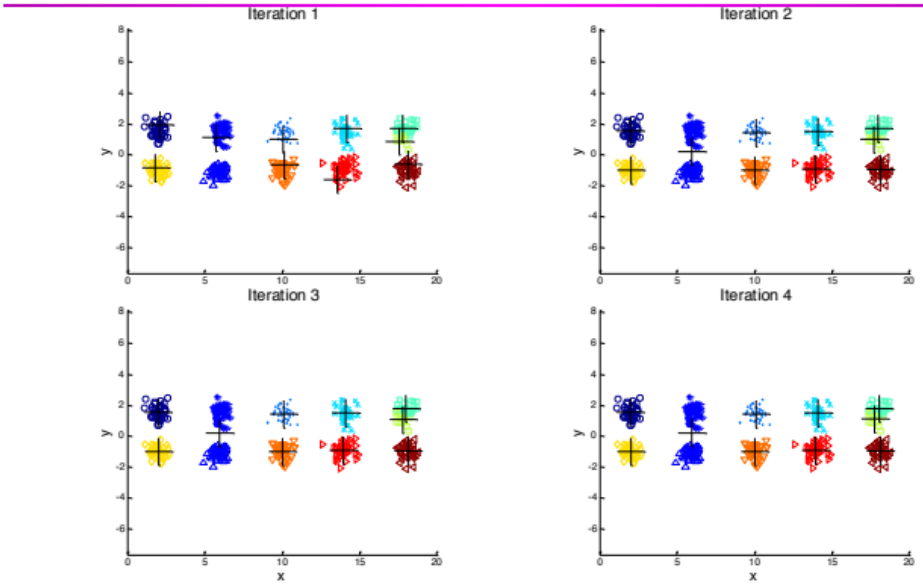


10 Clusters Example



Starting with some pairs of clusters having three initial centroids, while other have only one.

10 Clusters Example



Starting with some pairs of clusters having three initial centroids, while other have only one.

02/14/2018 数据挖掘导论，第 2 版 42

初始质心问题的解

●多次运行

有帮助，但可能性不在你这边

●采样并使用分层聚类来确定初始质心

●选择 k 个以上的初始质心，然后在这些初始质心中进行选择

选择分隔最广的

●后处理

●生成大量聚类，然后执行分层聚类●平分知识分子

不容易受到初始化问题的影响

02/14/2018 数据挖掘导论，第 2 版 43

Kmeans++

●这种方法可能比随机初始化慢，但在 SSE 方面，它始终能产生更好的结果

kmeans++算法保证了一个近似比率

o(对数 k)为期望值，其中 k 是中心数

●要选择一组初始质心，执行以下操作

1.随机选择一个初始点作为第一个质心 2。对于 k 1 步骤

3.对于 N 个点中的每一个，xi，1 ≤ i ≤ N，求最小平方

到当前选择的质心的距离，C1，…，Cj，1 ≤ j < k，

即最低 d 美元(xi)

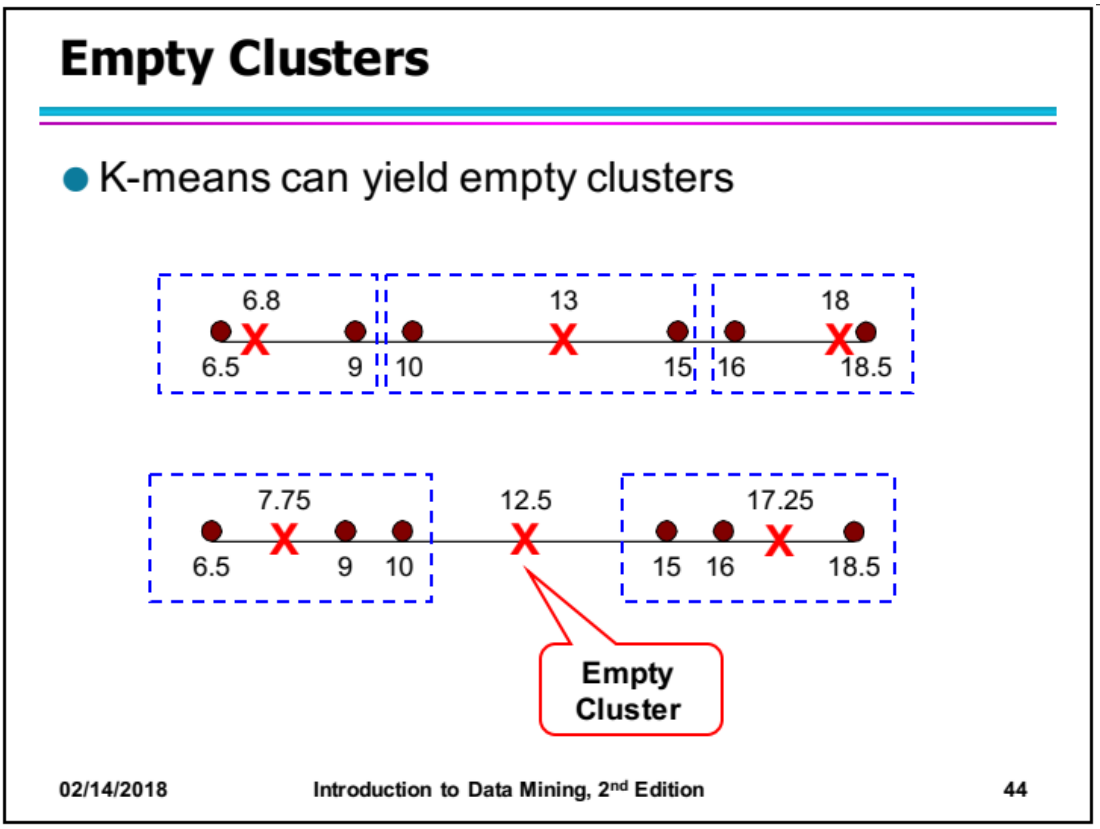
4.通过选择具有概率的点来随机选择新的质心

正比于

&'() d(Cjxi)

∑ &'() d(Cj，xi) +是

5.结束于



02/14/2018 数据挖掘导论，第 2 版 45
处理空集群
●基本的 Kmeans 算法会产生空簇
●几种策略
选择对上交所贡献最大的点
从具有最高 SSE 的集群中选择一个点
如果有几个空集群，上面可以是
重复几次。

02/14/2018 数据挖掘导论，第 2 版 46
增量更新中心
●在基本的 Kmeans 算法中，质心在所有点被分配到质心后被更新●另一种方法是在每次分配后更新质心(增量方法)
每个赋值更新零个或两个质心
更贵
引入订单相关性
永远不要得到空簇
可以用“重量”来改变影响

02/14/2018 数据挖掘导论，第 2 版 47
预处理和后处理
●预处理

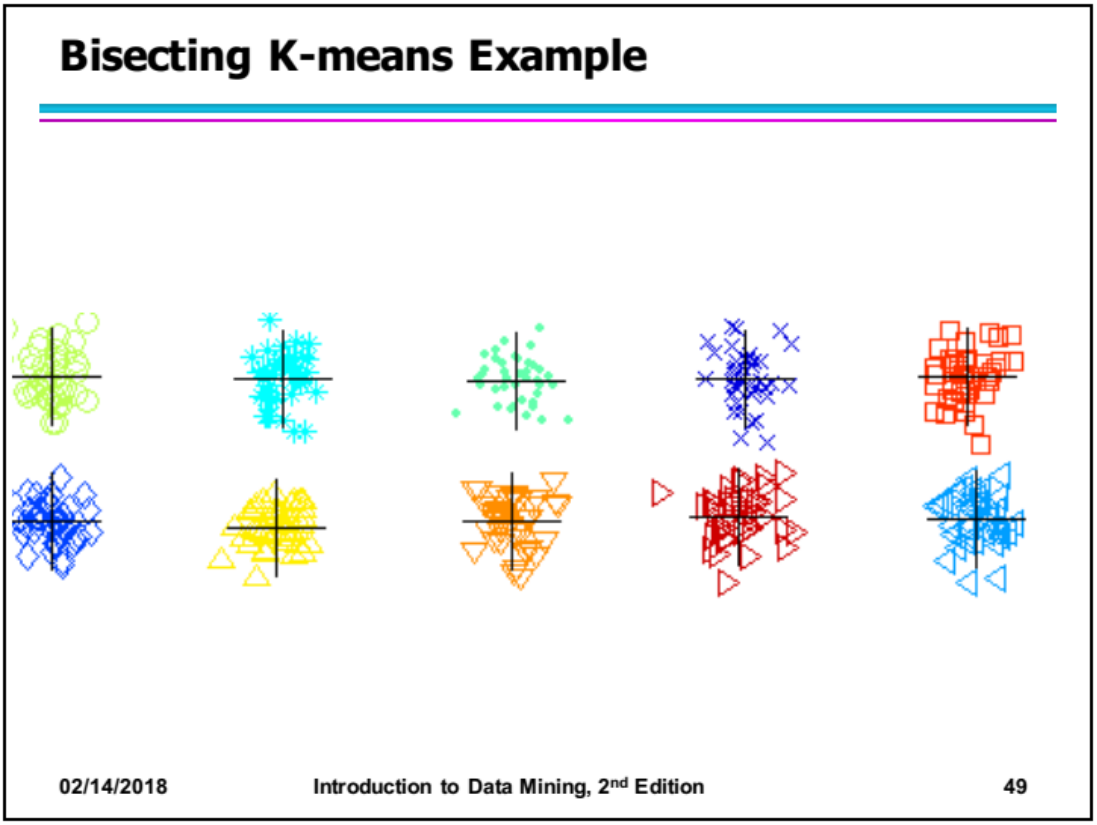
标准化数据
消除异常值
●后处理
消除可能代表异常值的小簇
拆分“松散”集群，即上交所相对较高的集群
合并“相近”且具有相对
低上交所
可以在群集过程中使用这些步骤
u ISODATA

Bisecting K-means

- Bisecting K-means algorithm
  - Variant of K-means that can produce a partitional or a hierarchical clustering

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

CLUTO: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview



02/14/2018 数据挖掘导论，第二版 50

分层聚类

●生成一组嵌套集群，这些集群被组织为

分层树

●可以可视化为树形图

一个树形图，记录了

合并或拆分

1 3 2 5 4 6

0.05

0.1

0.15

0.2

1

2

3

4

5

6

1

2

5

02/14/2018 数据挖掘导论，第 2 版 51

层次聚类的优势

●不必假设任何特定数量的

簇

可以通过以下方法获得任何期望数量的聚类

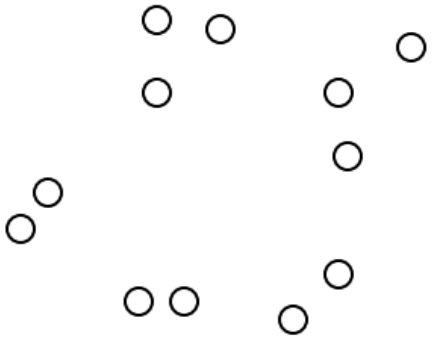
在适当的水平上切割树形图
●它们可能对应于有意义的分类法
生物科学中的例子(如动物界、系统发育重建等)

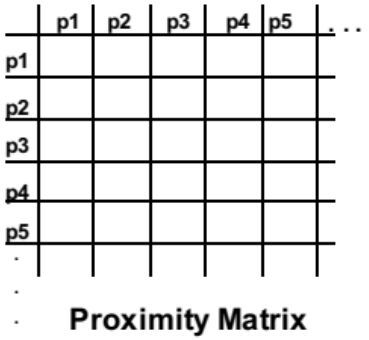
02/14/2018 数据挖掘导论，第 2 版 52
分层聚类
●两种主要类型的分层聚类
凝聚:
u 从点开始，作为单独的簇
u 在每个步骤中，合并最近的一对集群，直到只剩下一个集群(或 k 个集群)
分裂:
u 从一个包容的集群开始
u 在每个步骤中，分割一个集群，直到每个集群包含一个单独的点(或者有 k 个集群)
●传统的分层算法使用相似度或距离矩阵
一次合并或拆分一个集群

02/14/2018 数据挖掘导论，第 2 版 53
凝聚聚类算法
●最流行的分层聚类技术●基本算法很简单
1.计算邻近矩阵
2.让每个数据点成为一个集群
3.重复
4.合并两个最近的集群
5.更新邻近矩阵
6.直到只剩下一个集群
●关键操作是计算
两个集群
定义之间距离的不同方法
聚类区分不同的算法

Starting Situation

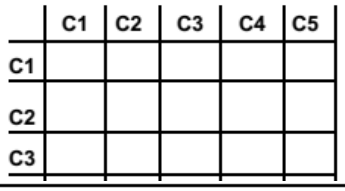
- Start with clusters of individual points and a proximity matrix

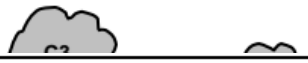




Intermediate Situation

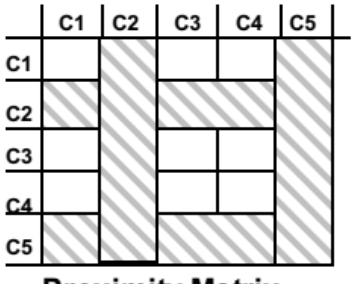
- After some merging steps, we have some clusters

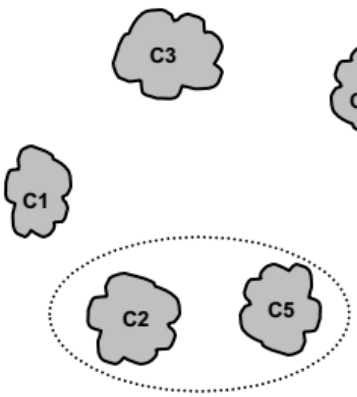


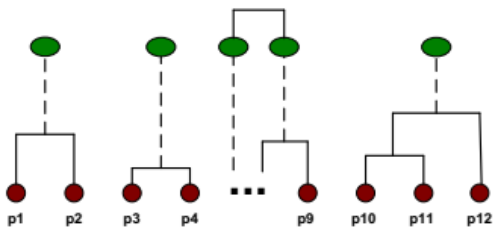


Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.

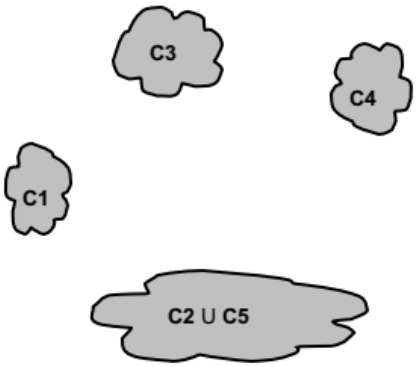






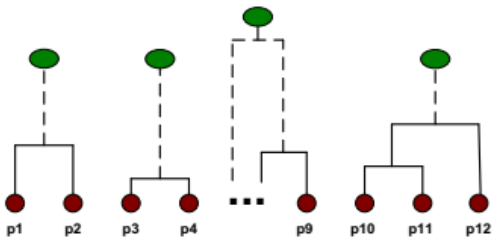
After Merging

- The question is “How do we update the proximity matrix?”

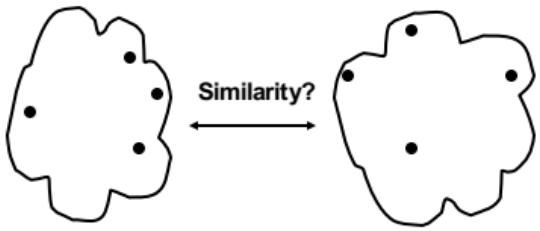


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix

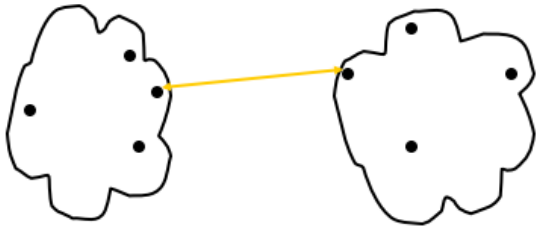


How to Define Inter-Cluster Distance



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						

How to Define Inter-Cluster Similarity

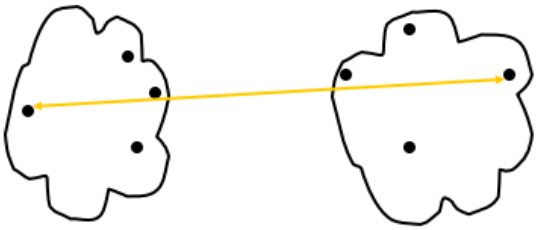


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward’s Method uses squared error

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

How to Define Inter-Cluster Similarity

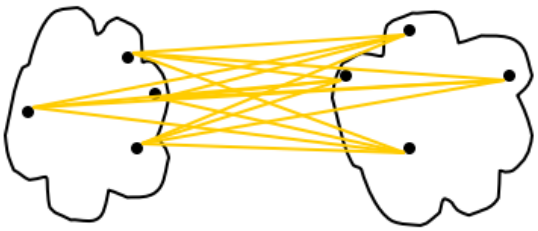


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

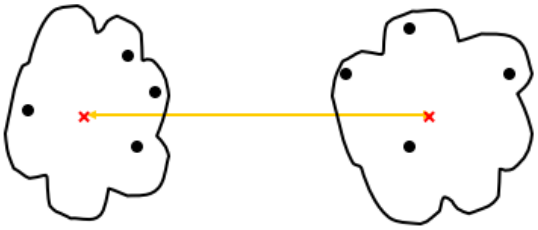
- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

02/14/2018 数据挖掘导论，第 2 版 63

最小或单链路

●两个集群的接近度基于两者

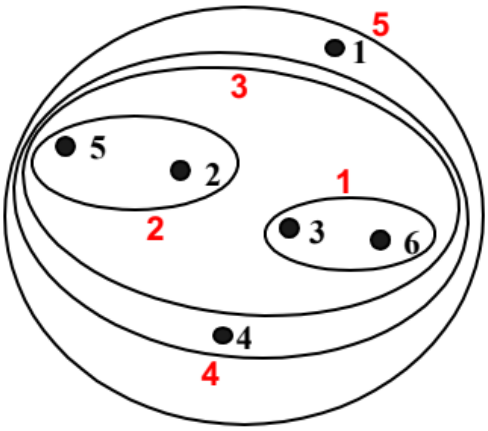
不同聚类中最近的点

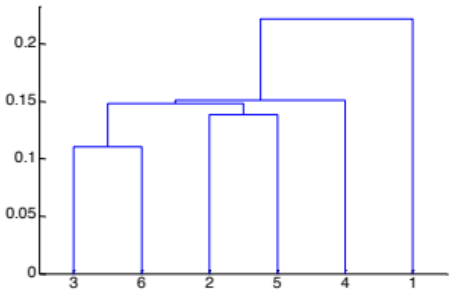
由一对点确定，即由邻近图中的一个链接确定

●示例:

距离矩阵:

Hierarchical Clustering: MIN





Nested Clusters

Dendrogram

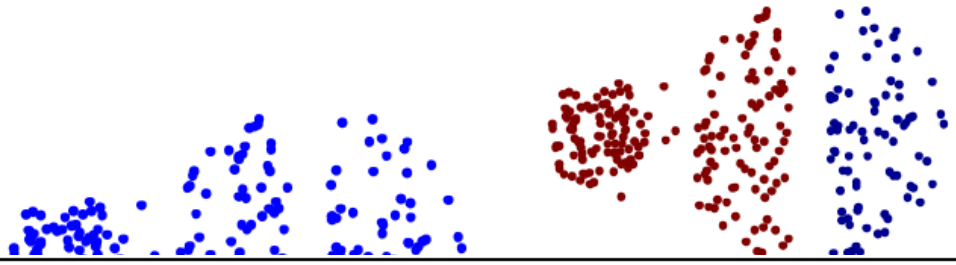
Strength of MIN

Original Points

Six Clusters

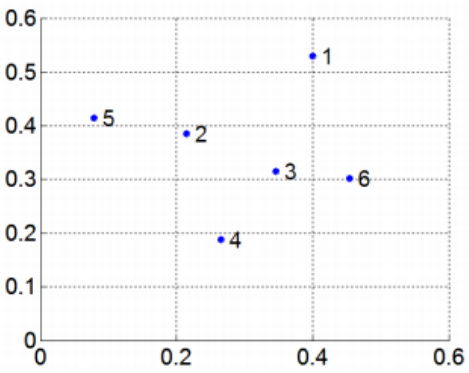
- Can handle non-elliptical shapes

Limitations of MIN



MAX or Complete Linkage

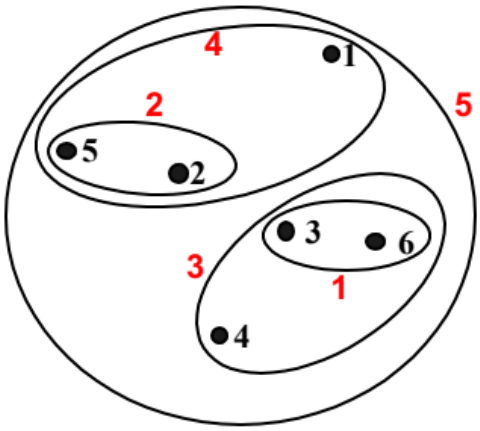
- Proximity of two clusters is based on the two most distant points in the different clusters
  - Determined by all pairs of points in the two clusters



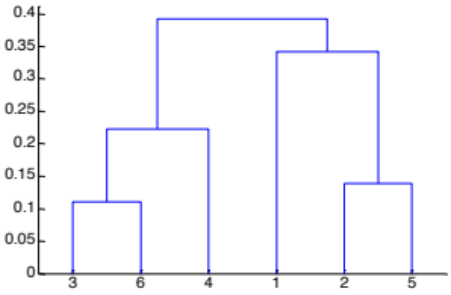
Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Hierarchical Clustering: MAX

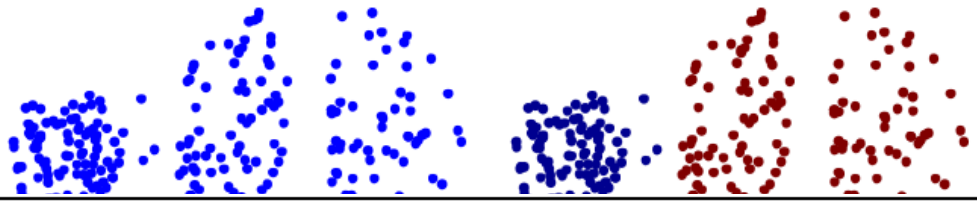


Nested Clusters

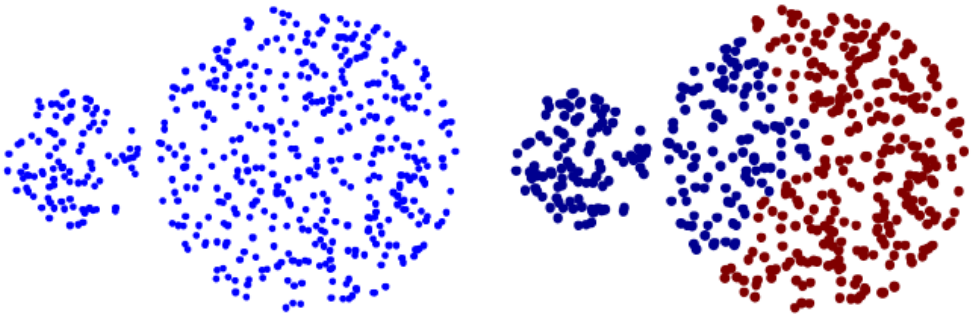


Dendrogram

Strength of MAX



Limitations of MAX



Original Points

Two Clusters

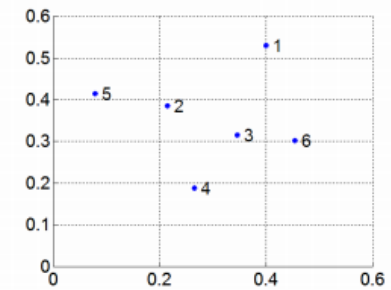
- Tends to break large clusters
- Biased towards globular clusters

Group Average

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{p_i \in \text{Cluster}_i, p_j \in \text{Cluster}_j} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| \times |\text{Cluster}_j|}$$

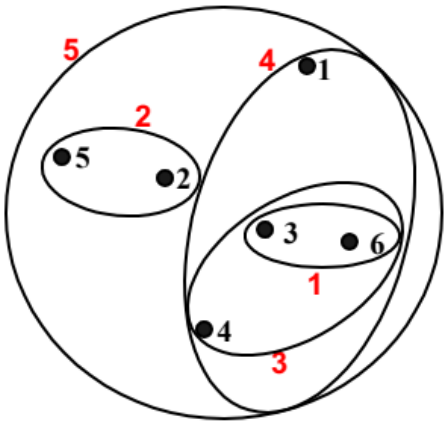
- Need to use average connectivity for scalability since total proximity favors large clusters



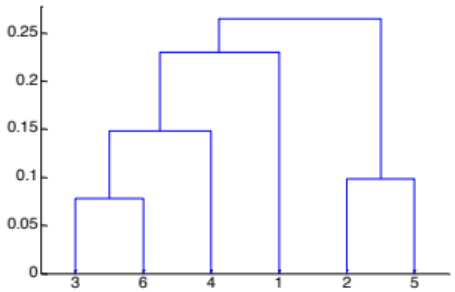
Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Hierarchical Clustering: Group Average



Nested Clusters

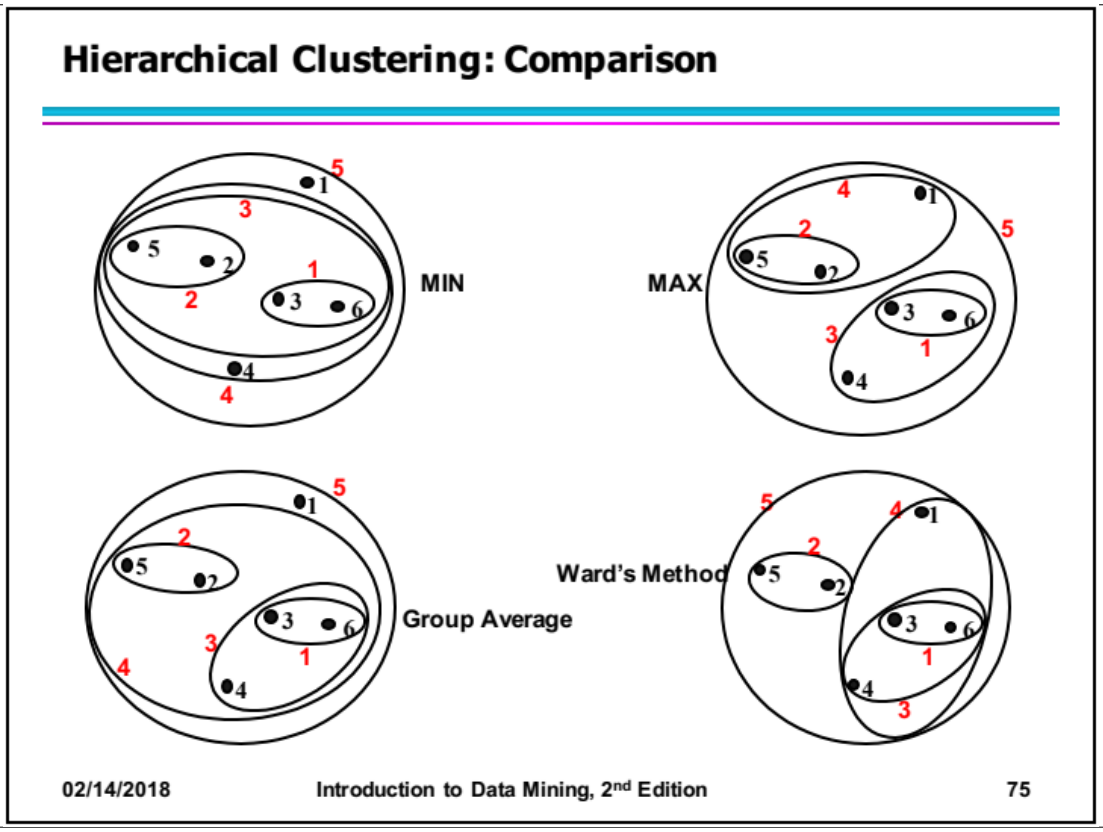


Dendrogram

02/14/2018 数据挖掘导论，第 2 版 73
分层聚类:组平均值
●单一链接和完整链接之间的折衷

●优势
不易受噪声和异常值的影响
●局限性
偏向球状星团

02/14/2018 数据挖掘导论，第 2 版 74
聚类相似性:沃德方法
●两个集群的相似性基于增加
当两个集群合并时的平方误差
如果点之间的距离为，则类似于组平均值
距离平方
●不易受噪声和异常值的影响
●偏向球状星团
●知识分子的等级模拟
可用于初始化 Kmeans



02/14/2018 数据挖掘导论，第 2 版 76
分裂的层次聚类
●构建最小生成树
从由任意点组成的树开始
在连续的步骤中，寻找最近的一对点(p，q)，使得一个点(p)在当前树中，而另一个点(q)不在树中，将 q 添加到树中，并在 p 和 q 之间放置一条边

MST: Divisive Hierarchical Clustering

- Use MST for constructing hierarchy of clusters

Algorithm 7.5 MST Divisive Hierarchical Clustering Algorithm

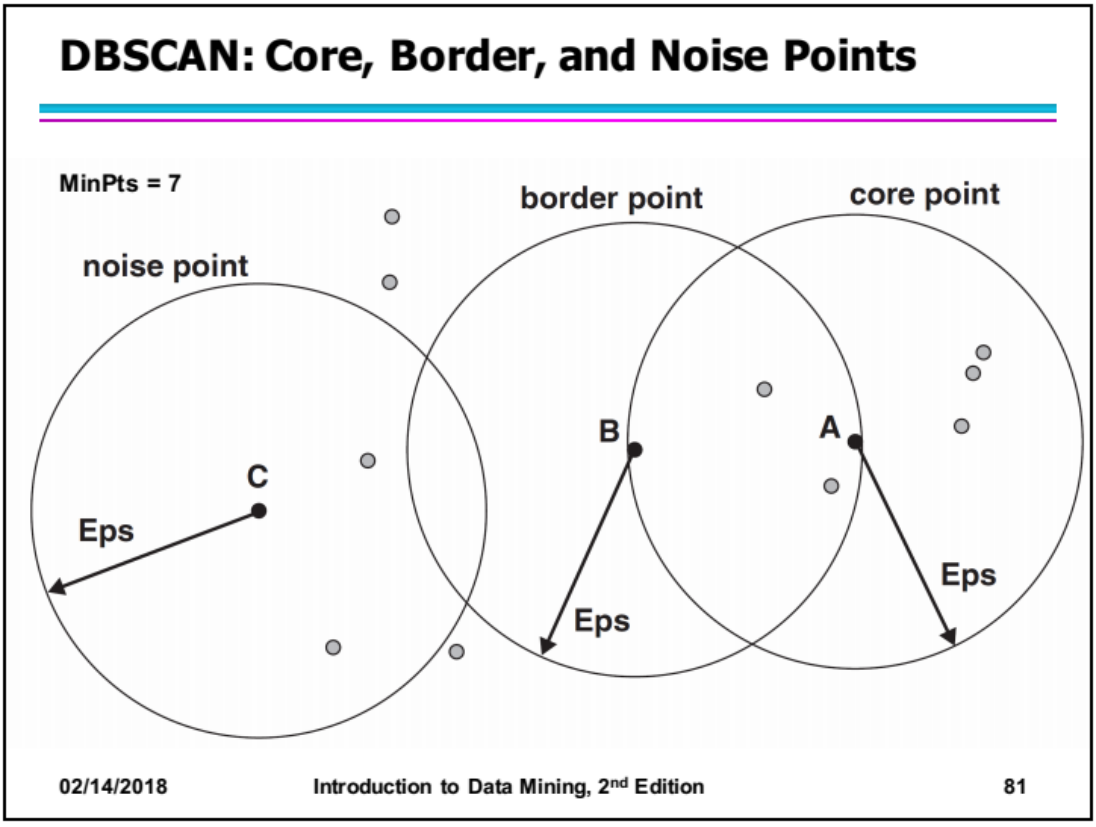
1: Compute a minimum spanning tree for the proximity graph.
2: **repeat**
3: Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).
4: **until** Only singleton clusters remain

02/14/2018 数据挖掘导论，第 2 版 78
分层聚类:时间和空间要求
●0(N)空间，因为它使用邻近矩阵。
n 是点数。
●在许多情况下没有时间
有 N 个步骤，每个步骤的大小为 N，
必须更新和搜索邻近矩阵
只要聪明一点，复杂性可以减少到 0(N)

02/14/2018 数据挖掘导论，第 2 版 79
分层聚类:问题与局限
●一旦决定合并两个集群，
这是无法挽回的
●没有直接最小化的全局目标函数●不同的方案有一个或多个问题
更多以下内容:
对噪声和异常值的敏感性
难以处理不同大小和非球形的星团
打破大型集群

02/14/2018 数据挖掘导论，第 2 版 80
DBSCAN
● DBSCAN 是一种基于密度的算法。
密度=指定半径内的点数(Eps)
一个点是一个核心点，如果它至少有指定数量的

Eps 内的点数(MinPts)
这些点位于一个星团的内部
u 计算点本身
边界点不是核心点，而是在邻域内
核心点
噪声点是任何不是核心点或边界的点
要点

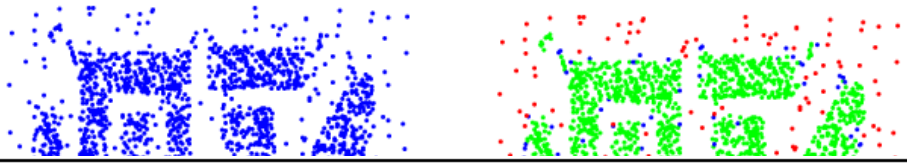


DBSCAN Algorithm

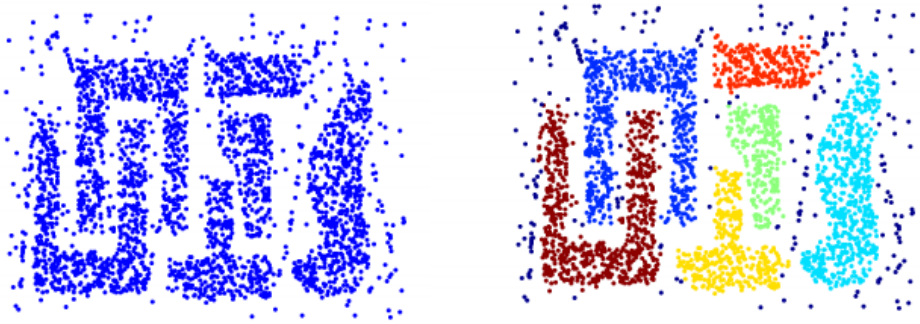
- Eliminate noise points
- Perform clustering on the remaining points

```
current_cluster_label ← 1
for all core points do
  if the core point has no cluster label then
    current_cluster_label ← current_cluster_label + 1
    Label the current core point with cluster label current_cluster_label
  end if
  for all points in the Eps-neighborhood, except i^th the point itself do
    if the point does not have a cluster label then
      Label the point with cluster label current_cluster_label
    end if
  end for
end for
```


DBSCAN: Core, Border and Noise Points



When DBSCAN Works Well

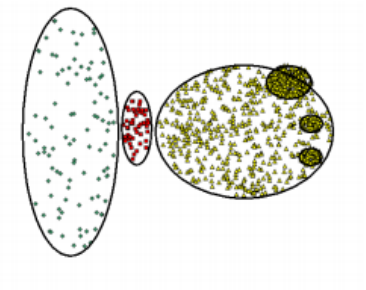


Original Points

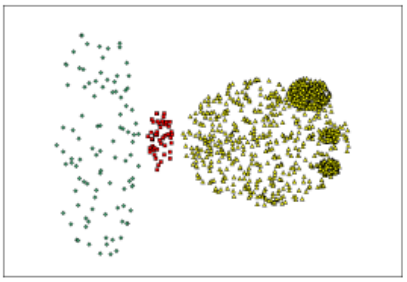
Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

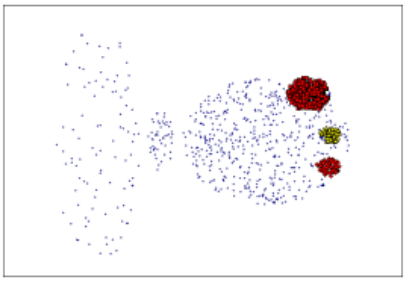
When DBSCAN Does NOT Work Well



Original Points



(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

- Varying densities
- High-dimensional data

02/14/2018 数据挖掘导论，第 2 版 86

DBSCAN:确定每股收益和最小交易点

●想法是，对于一个集群中的点，它们的第 k 个最近邻居之间的距离大致相同

●噪声点的第 k 个最近邻居在更远的地方

距离

●所以，画出每一点到它的 kth 的排序距离

最近邻

02/14/2018 数据挖掘导论，第 2 版 87

聚类有效性

●对于监督分类，我们有多种方法来评估我们的模型有多好

准确性、精确性、召回率

●对于聚类分析，类似的问题是如何评估结果聚类的“良好性”？

●但是“集群在旁观者的眼里”！

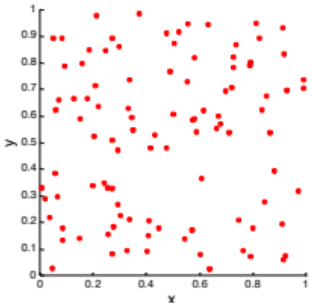
●那我们为什么要评估它们?

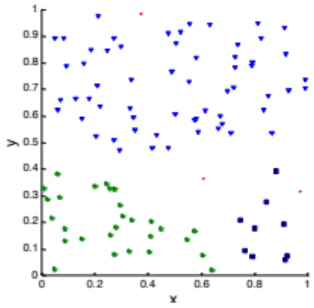
为了避免在噪音中发现模式

比较聚类算法比较两组聚类比较两组聚类

Clusters found in Random Data

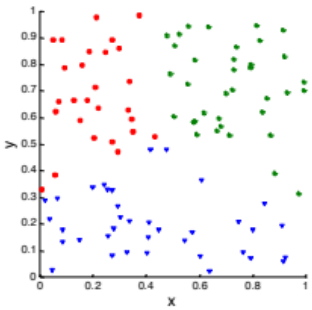
Random Points

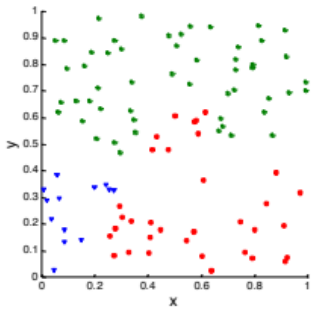




DBSCAN

K-means





Complete Link

02/14/2018 Introduction to Data Mining, 2nd Edition 88

02/14/2018 数据挖掘导论，第 2 版 89

1.确定一组数据的聚类趋势，即区分数据中是否实际存在非随机结构。

2.将聚类分析的结果与外部已知的结果进行比较

结果，例如外部给定的类别标签。

3.评估聚类分析的结果与数据的吻合程度

不参考外部信息。

仅使用数据

4.将两组不同的聚类分析结果进行比较

确定哪个更好。

5.确定群集的“正确”数量。

对于 2、3 和 4，我们可以进一步区分我们是否想要

评估整个集群或单个集群。

聚类验证的不同方面

02/14/2018 数据挖掘导论，第 2 版 90

●用于判断各个方面的数值度量

聚类有效性分为以下三种类型。

外部索引:用于衡量集群标签的范围

匹配外部提供的类别标签。

熵

内部指数:用来衡量一个聚类的好坏

不考虑外部信息的结构。

误差平方和

相对指数:用于比较两个不同的聚类或

集群。
u 该函数通常使用外部或内部指数，如上证综指或熵
●有时这些被称为标准，而不是指数
然而，有时标准是总体策略，而指数是
实现标准的数值度量。
聚类有效性的度量

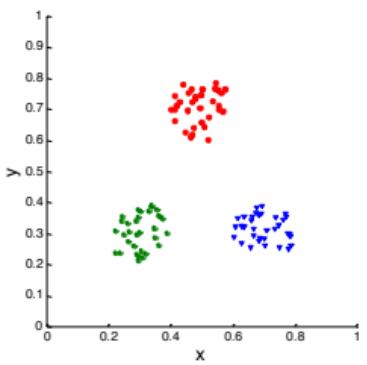
02/14/2018 数据挖掘导论，第 2 版 91
●两个矩阵
邻近矩阵
理想相似矩阵
u 每个数据点一行一列
如果关联的点对属于同一个聚类，则条目为 1
如果关联的点对属于不同的聚类，则条目为 0
●计算两个矩阵之间的相关性
因为矩阵是对称的，所以只有
需要计算 n(n1) / 2 个条目。
●高相关性表示属于同一聚类的点彼此靠近。
●对于一些基于密度或邻近性的指标来说，这不是一个好的衡量标准
集群。
通过相关性测量聚类有效性

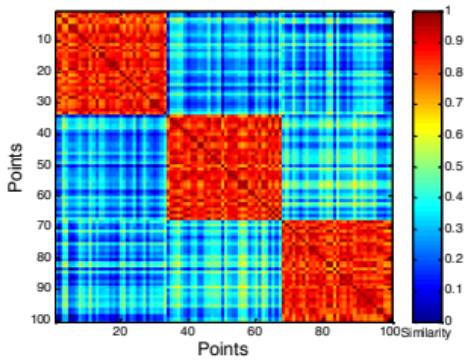
Measuring Cluster Validity Via Correlation

- Correlation of ideal similarity and proximity matrices for the K-means clusterings of the following two data sets.

Using Similarity Matrix for Cluster Validation

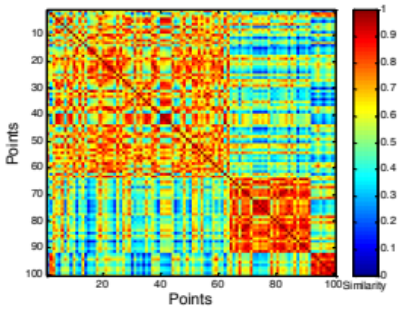
- Order the similarity matrix with respect to cluster labels and inspect visually.

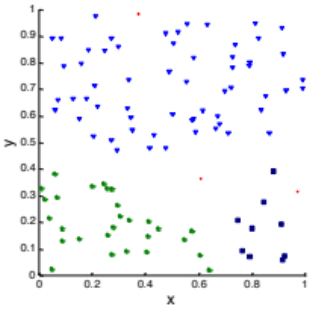




Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp

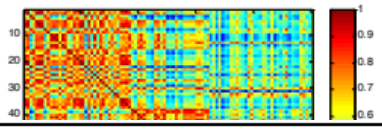


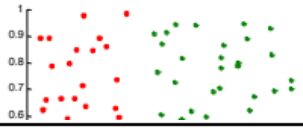


DBSCAN

Using Similarity Matrix for Cluster Validation

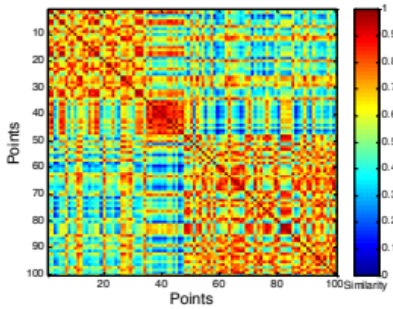
- Clusters in random data are not so crisp

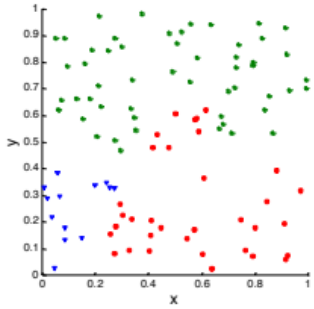




Using Similarity Matrix for Cluster Validation

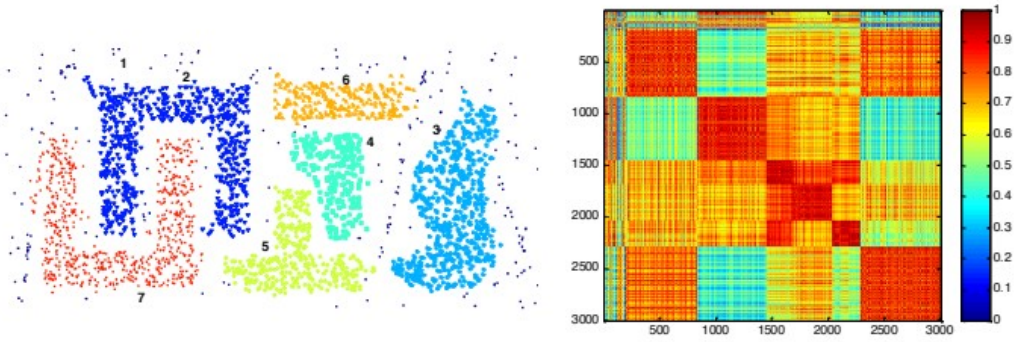
- Clusters in random data are not so crisp





Complete Link

Using Similarity Matrix for Cluster Validation



DBSCAN

02/14/2018 数据挖掘导论，第 2 版 98

●更复杂图形中的簇没有很好地分开

●内部指数:用于衡量一个聚类的优劣

不考虑外部信息的结构

南东南

● SSE 非常适合比较两个集群或两个集群

(平均上证指数)。

●也可用于估计集群的数量

内部措施:上交所

2 5 10 15 20 25 30

1

2

3

4

5

6

7

8

9

10

K

南东南

5 10 15

-6

-4

-2

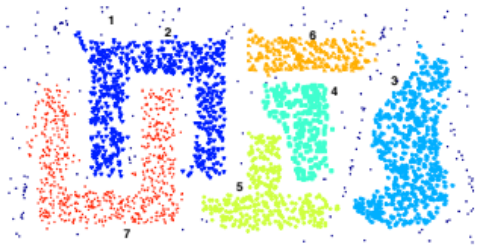
0

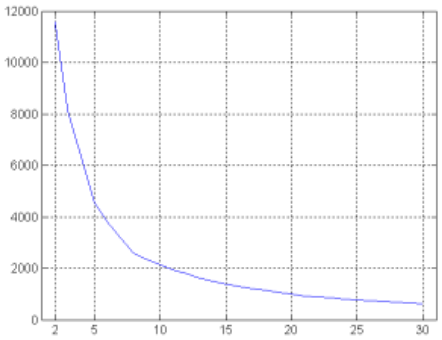
2

4
6

Internal Measures: SSE

- SSE curve for a more complicated data set



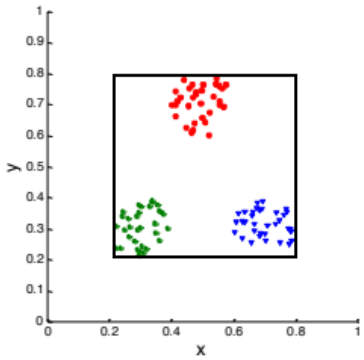


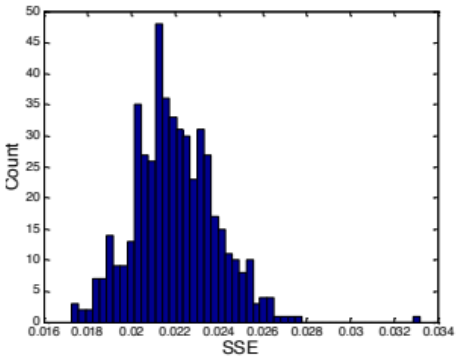
SSE of clusters found using K-means

02/14/2018 数据挖掘导论，第 2 版 100
●需要一个框架来解释任何衡量标准。
例如，如果我们的评估度量值为 10，那么
好，公平，还是贫穷？
●统计数据为集群有效性提供了一个框架
聚类结果越“不典型”，就越有可能代表
数据中的有效结构
可以比较由随机数据或
聚类结果的聚类。
u 如果索引值不太可能，则聚类结果有效
这些方法更复杂，也更难理解。
●用于比较两组不同聚类的结果
分析，框架是不必要的。
然而，存在着两者之间的区别
指数值很重要
聚类有效性框架

Statistical Framework for SSE

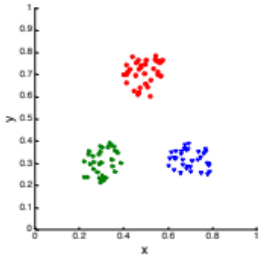
- Example
  - Compare SSE of 0.005 against three clusters in random data
  - Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values

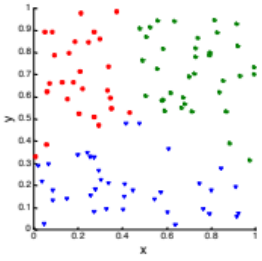


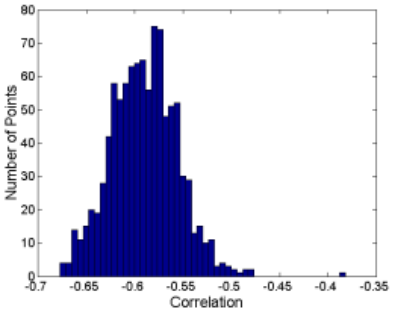


Statistical Framework for Correlation

- Correlation of ideal similarity and proximity matrices for the K-means clusterings of the following two data sets.







Corr = -0.9235 Corr = -0.5810

02/14/2018 数据挖掘导论，第 2 版 103

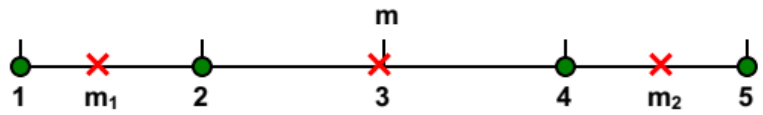
●集群内聚力:衡量集群中对象之间的紧密程度

例如:上交所

●集群分离:衡量不同或良好程度
将一个集群与其他集群分开
●示例:平方误差
内聚力通过聚类内平方和来衡量
间隔是通过两个聚类之间的平方和来衡量的
其中|Ci|是群集 I 的大小
内部衡量:凝聚力和分离
∑2
∈
= = I x C I I SSE WSS x m()
=∑-I BSS Ci m mi()

Internal Measures: Cohesion and Separation

- Example: SSE
  - BSS + WSS = constant



K=1 cluster:

$$SSE = WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

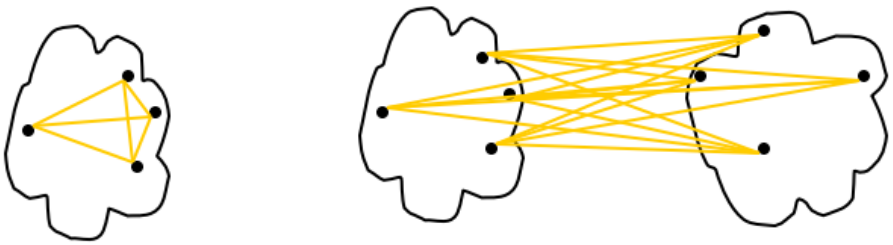
$$SSE = WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Internal Measures: Cohesion and Separation

- A proximity graph based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion separation

02/14/2018 数据挖掘导论，第 2 版 106
●轮廓系数结合了内聚和分离的思想，但对于单个点，以及聚类和集群
计算 a =到其聚类中的点的平均距离
计算 b = min(I 到另一个聚类中的点的平均距离)
点的轮廓系数由 s = (b a) /最大值(a，b)给出
通常在 0 和 1 之间。
越接近 1 越好。
●可以计算一个聚类或一个聚类的平均轮廓系数
内部测量:轮廓系数
使用的距离
计算一个
i
使用的距离
计算 b

External Measures of Cluster Validity: Entropy and Purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

02/14/2018 数据挖掘导论，第 2 版 108
“聚类结构的验证是最
聚类分析中困难和令人沮丧的部分。如果没有这方面的努力，聚类分析仍将是一门黑色艺术，只有那些有经验和勇气的真正信徒才能接触到。"
数据聚类算法，贾恩和杜贝斯
集群有效性的最终评价