数据库系统概论新技术篇

数据挖掘

李翠平

中国人民大学信息学院

概　览

- 什么是聚类？
- 聚类的主要过程
- 如何对聚类结果进行评价？
- 常用的聚类方法有哪些？
- 一种典型的聚类方法：K-Means聚类

什么是聚类

- 属于模型挖掘，更确切地讲，是描述建模的过程
- 根据数据特征找出数据间的相似性，将相似的数据分成一个类
  - 聚在同一个类中的数据要足够相似
  - 没在同一个类中的数据要尽量不相似
- 无监督学习：没有预设的类标号

聚类应用的领域

- 作为一个独立的工具对数据分布进行分析
- 可以作为其他算法（如分类等）的预处理步骤
- 模式识别
- 空间数据分析
- 图像处理
- 经济科学(尤其是市场研究)
- WWW

聚类应用的例子

- 市场分析：帮助市场分析人员从客户基本库中发现不同的客户群，并且用购买模式来刻画不同的客户群的特征
- 土地使用：在地球观测数据库中用以确定相似地区
- 城市规划：根据房子的类型，价值，和地理位置对一个城市中的房屋分组

聚类方法的好坏

- 好的聚类方法能产生高质量的聚类。所谓高质量，指：
  - 类中的对象高度相似
  - 类间的对象高度不相似
- 聚类的质量与什么有关？
  - 相似性度量及其实现方法
- 聚类方法的好坏也可以按照它是否能够发现更多的隐含模式来度量

好的聚类算法应该具有的特征

- 可伸缩性
- 能够处理各种不同类型的属性
- 能够发现任意形状的聚类
- 在决定输入参数的时候，对领域知识的需求要小
- 能够处理噪声和异常点
- 对输入数据的顺序不敏感
- 可以处理高维数据
- 可以和用户制定的限定条件相结合
- 可解释性和使用性好

主要的聚类算法

1、基于划分的方法

给定一个n个对象或元组的数据库，划分方法构建数据的k个划分，每个划分表示一个聚类，并且k<=n。

也就是说，它将数据划分为k个组，同时满足如下的要求：

（1）每个组至少包含一个对象；

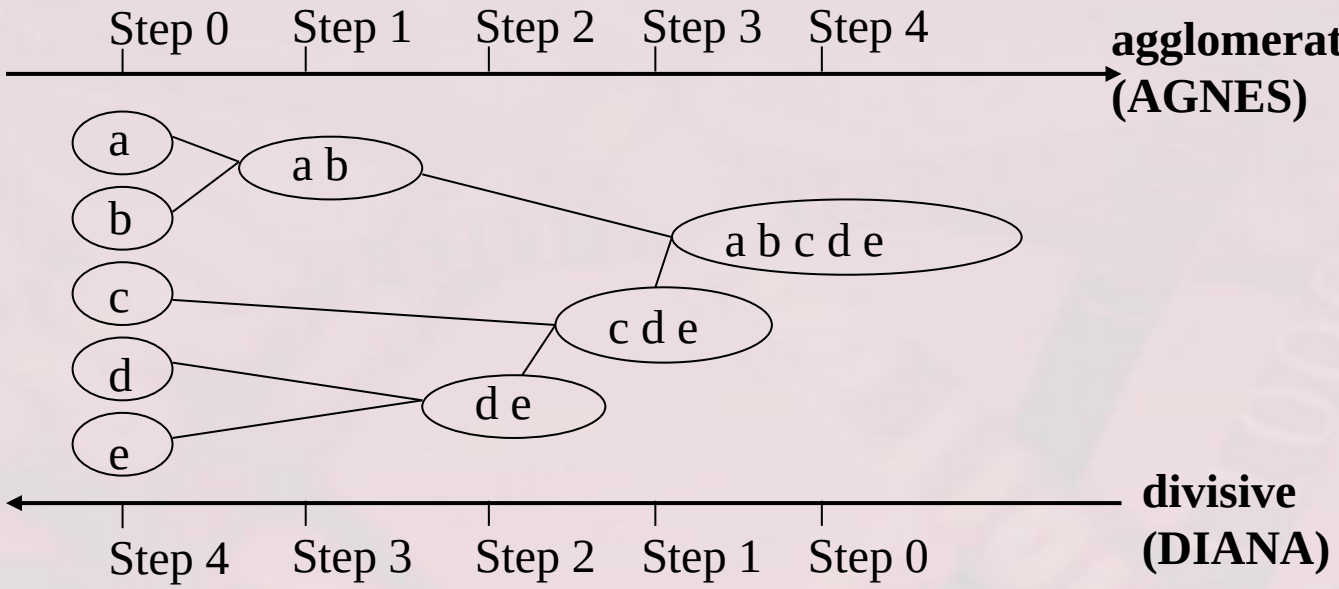
（2）每个对象必须属于且只属于一个组。

2、基于层次的聚类方法

主要思想是把数据对象排列成一个聚类树，在需要的层次上对其进行切割，相关联的部分构成一个cluster。基于层次的聚类方法有两种类型：

（1）聚合层次聚类。最初每个对象是一个cluster，然后根据它们之间的相似性，对这些原子的cluster进行合并。大多数层次方法属于这一类，它们的主要区别是cluster之间的相似性的定义不同

（2）划分层次聚类，它与上面的过程正好相反


Step 0
Step 1
Step 2
Step 3
Step 4
agglomerative
(AGNES)
a
b
a b
a b c d e
c
c d e
d
d e
e
divisive
(DIANA)
Step 4
Step 3
Step 2
Step 1
Step 0

- 用户可以指定算法终止的条件，例如，聚类的个数或每个cluster的半径低于某个阈值。
- 弱点在于合并或分裂点的选取问题，因为一组对象一旦合并或分裂，就不能有undo的操作
- 时间复杂度为O（N^2），对于处理大数据量有性能问题。

3、基于密度的方法

绝大多数划分方法基于对象之间的距离进行聚类。这样的方法只能发现凸状的簇，而在发现任意形状的簇上遇到了困难

基于密度的聚类方法的主要思想是：只要临近区域的密度（对象或数据点的数目）超过某个阈值，就继续聚类。也就是说，对给定类中的每个数据点，在一个给定范围的区域中必须包含至少某个数目的点。这样的方法可以用来过滤“噪音”数据，发现任意形状的簇。

4、基于方格的方法

- 把多维数据空间划分成一定数目的单元，然后在这种数据结构上进行聚类操作。
- 该类方法的特点是它的处理速度，因为其速度与数据对象的个数无关，而只依赖于数据空间中每个维上单元的个数。

5、基于模型的方法

（1）神经网络方法

（2）统计的方法

K-means算法

K-Means方法是MacQueen1967年提出的

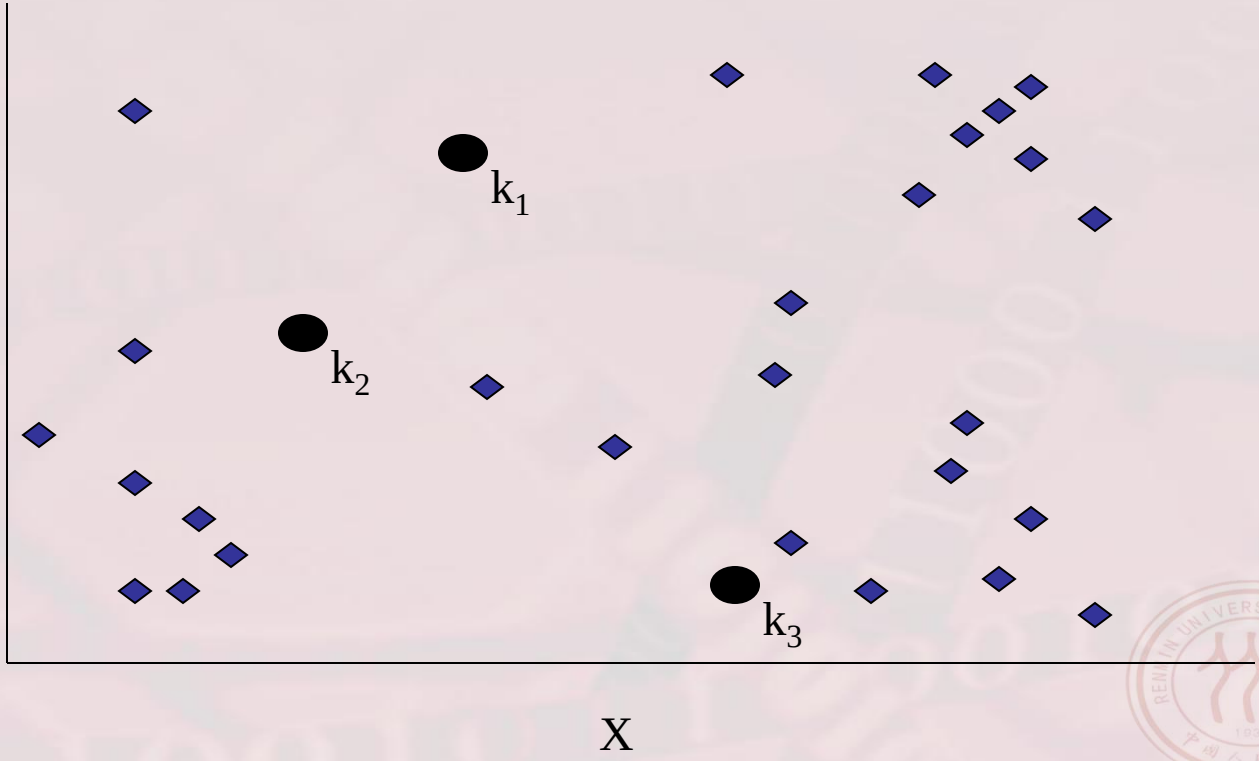
给定一个数据集合X和一个整数K（≤n），K-Means方法是将X分成K个聚类并使得在每个聚类中所有值与该聚类中心距离的总和最小

K-Means聚类方法分为以下几步：

[1] 给K个cluster选择最初的中心点，称为K个Means
[2] 计算每个对象和每个中心点之间的距离
[3] 把每个对象分配给距它最近的中心点所属的cluster
[4] 重新计算每个cluster的中心点
[5] 重复2，3，4步，直到算法收敛

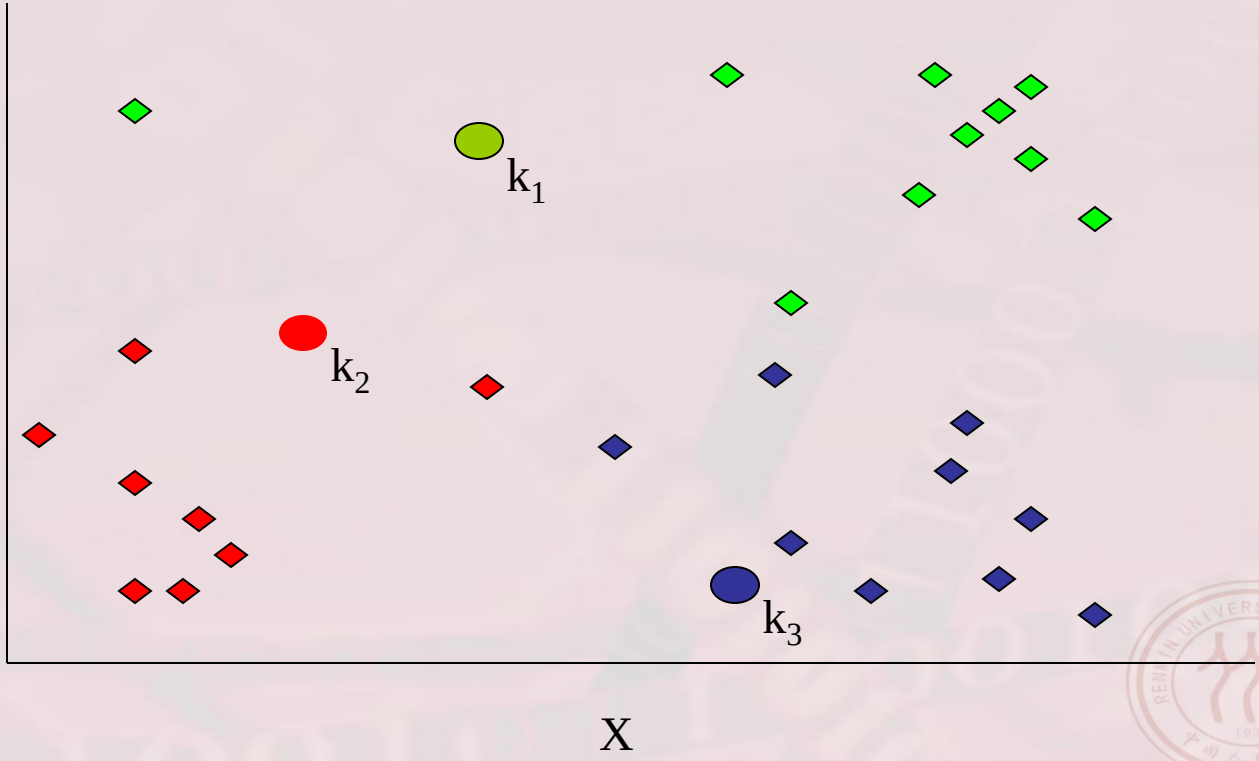
k-Means Example (I)

Pick 3 initial cluster centers (randomly)



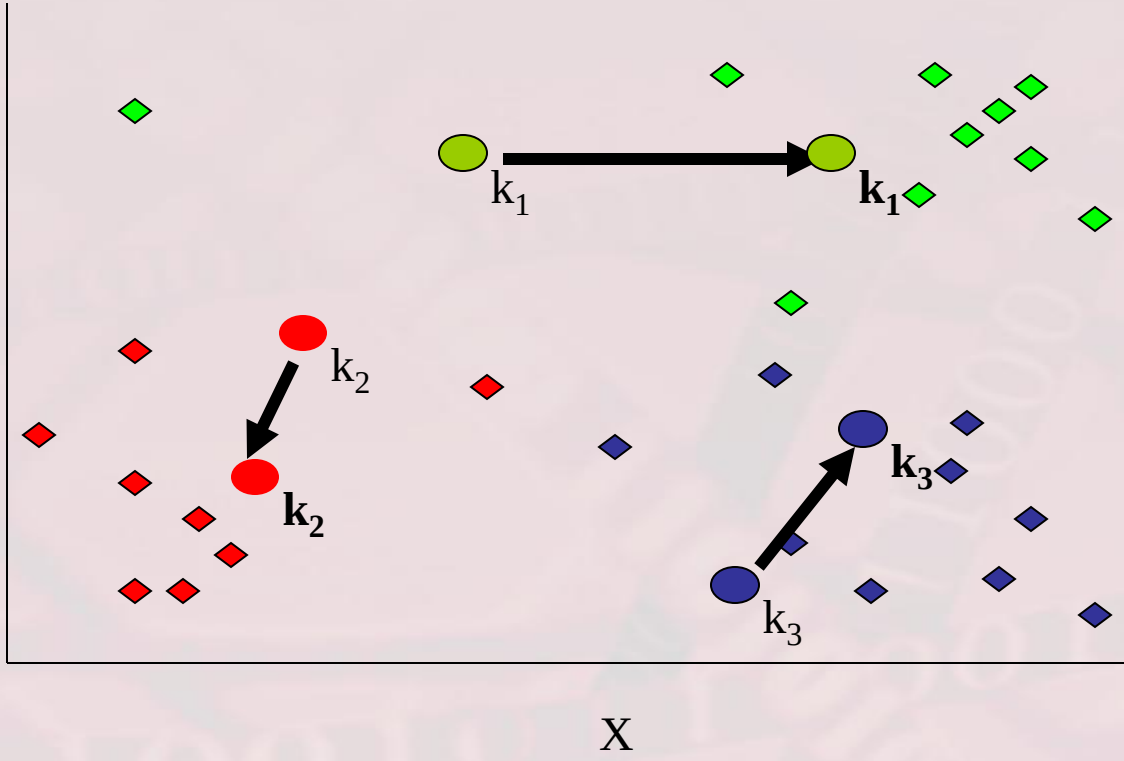
k-Means Example (II)

Assign each point to the closest cluster center



k-Means Example (III)

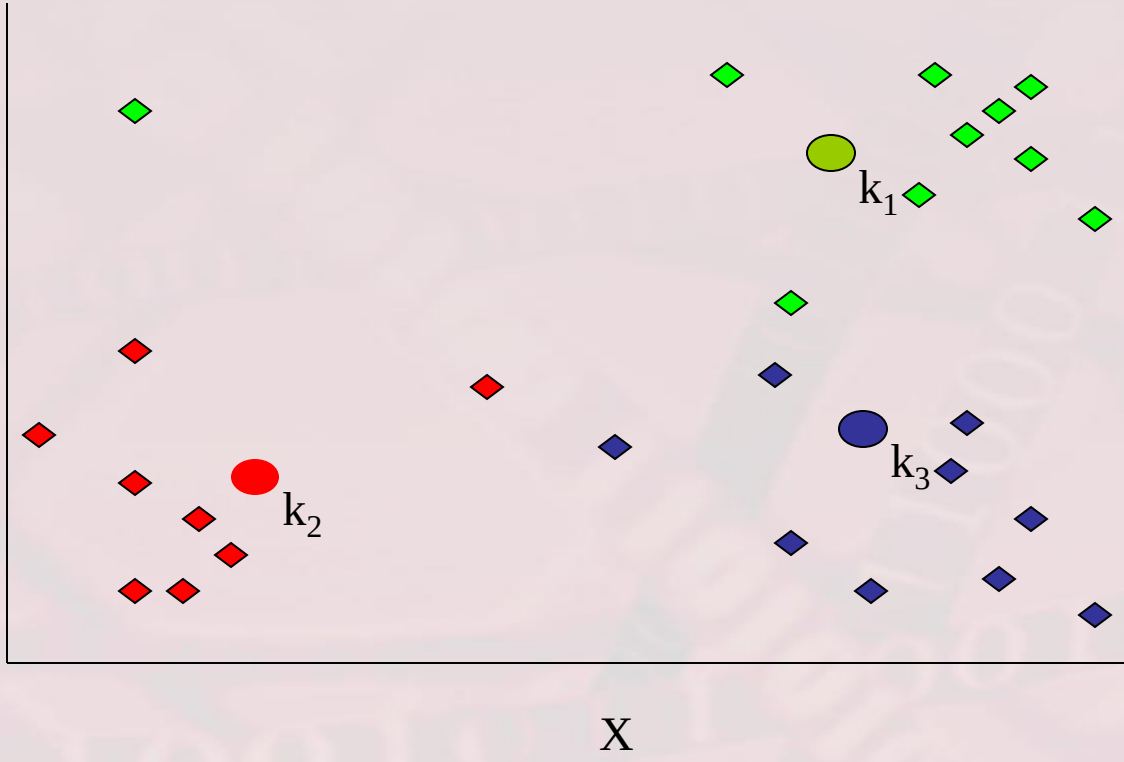
Move each cluster center to the mean of each cluster



k-Means Example (IV)

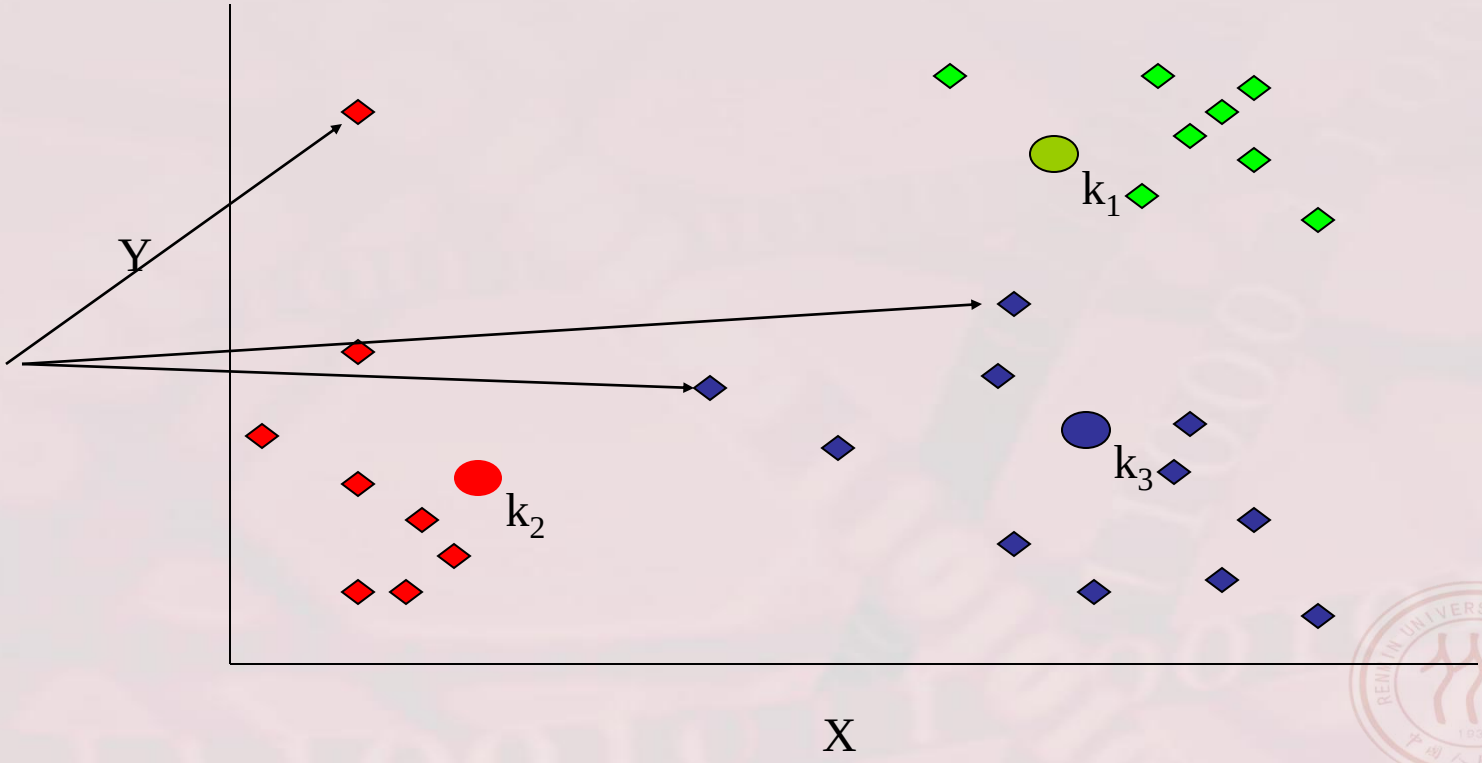
Reassign points closest to a different new cluster center

Q: Which points are reassigned?



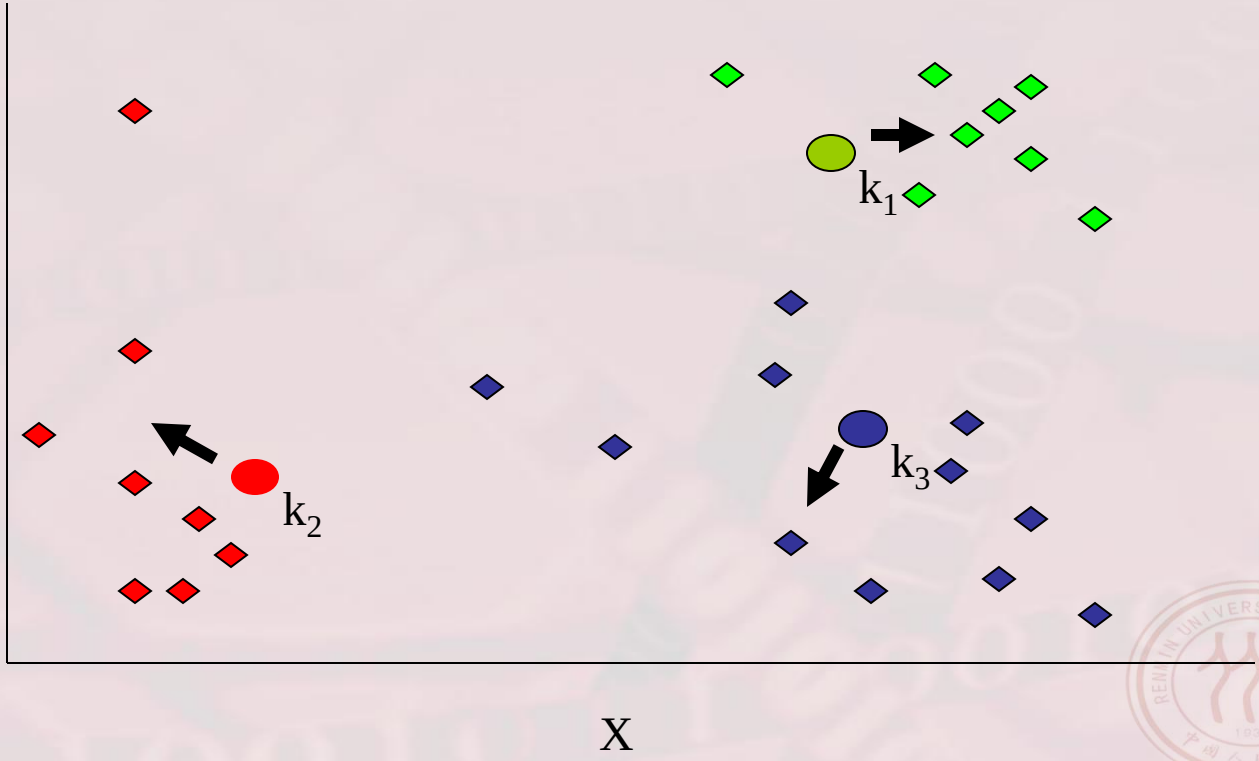
k-Means Example (V)

A: three points with animation



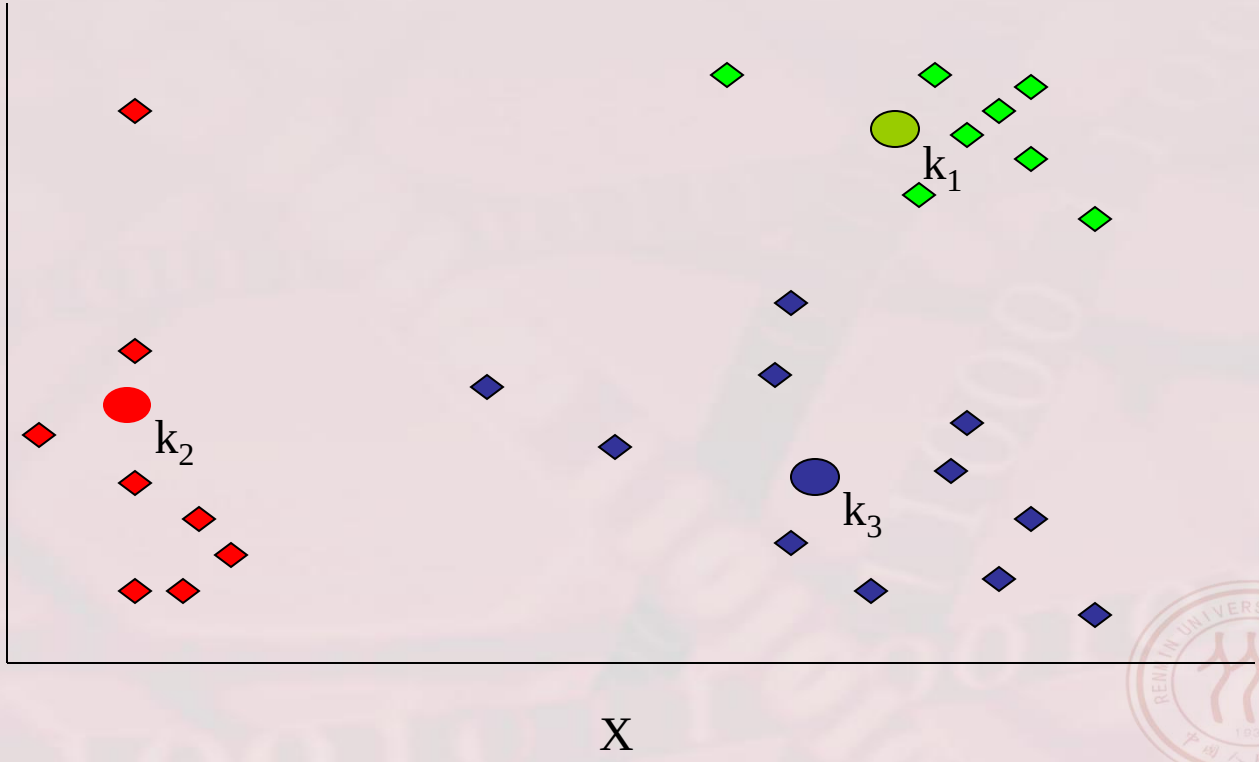
k-Means Example (VI)

re-compute
cluster means



k-Means Example (VII)

move cluster centers to cluster means



讨论

- 如何选择种子 C^0
- 可能的方法包括：
  - 随机的选择 C^0
  - 使用某种聚类算法的结果作为 C^0
  - *K*-means 采用的是hill-climbing，只找到局部最优，结果依赖于C^0

❖ K-Means方法具有下面的优点：

（1）对于处理大数据量具有可扩充性和高效率。算法的复杂度是O（tkn），其中n是对象的个数，k是cluster的个数，t是循环的次数，通常k，t<<n

（2）可以实现局部最优化,如果要找全局最优，可以用退火算法或者遗传算法

❖ K-Means方法也有以下缺点：

（1）Cluster的个数必须事先确定，在有些应用中，事先并不知道cluster的个数。

（2）K个中心点必须事先预定，而对于有些字符属性，很难确定中心点。

（3）不能处理噪音数据。

（4）不能处理有些分布的数据（例如凹形）

❖K-Means方法的变种

(1) K-Modes ：处理分类属性

(2) K-Prototypes：处理分类和数值属性

它们与K-Means方法的主要区别在于：

（1）最初的K个中心点的选择不同。

（2）距离的计算方式不同。

（3）计算cluster的中心点的策略不同。

谢 谢！