HUAWEI

7.1 机器学习中的聚类问题

目录

Contents

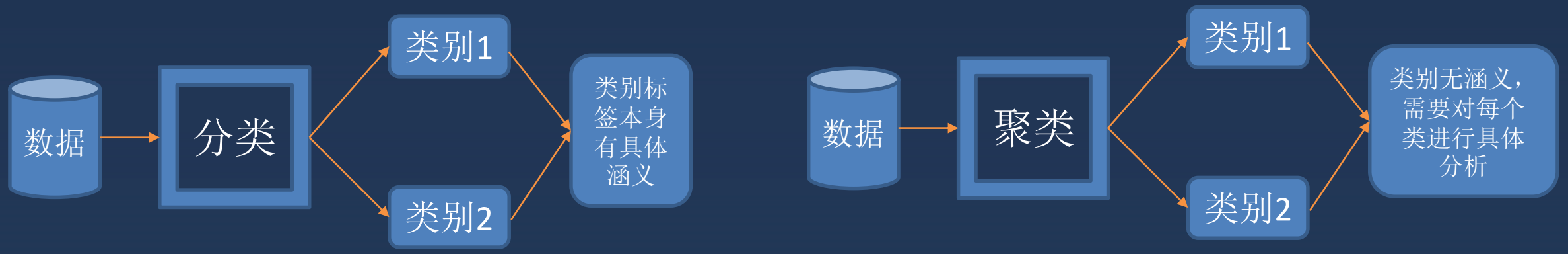
聚类

聚类（ Clustering ）就是按照某个特定标准把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。即聚类后同一类的数据尽可能聚集到一起，不同数据尽量分离。

聚类和分类的区别：

1、分类 (Classification)，对于一个分类器，通常需要你告诉它“这个东西被分为某某类”这样一些例子，理想情况下，一个分类器会从它得到的训练集中进行“学习”，从而具备对未知数据进行分类的能力，这种提供训练数据的过程通常叫做监督学习 (supervised learning)。

2、聚类算法通常又被称为无监督学习，因为与监督学习不同，在聚类中那些表示数据类别的分类或者分组信息是没有的。 简单地说就是把相似的东西分到一组，聚类的时候，我们并不关心某一类是什么，我们需要实现的目标只是把相似的东西聚到一起。因此，一个聚类算法通常只需要知道如何计算相似度就可以开始工作了，因此聚类通常并不需要使用训练数据进行学习，这在机器学习中被称作无监督学习 (unsupervised learning)。



目录

Contents

划分方法

定义：给定一个n个对象的集合，将其按照一定的规则划分为k个分区，其中每个分区表示一个簇，k<=n，每个分区至少包含一个对象。基本划分方法是互斥的。

使用迭代的重定位技术：

1、设定规则。
2、给定一个初始划分。
3、根据规则计算这个划分包含的对象集。
4、更新划分。
5、循环3和4，直到无法更新划分

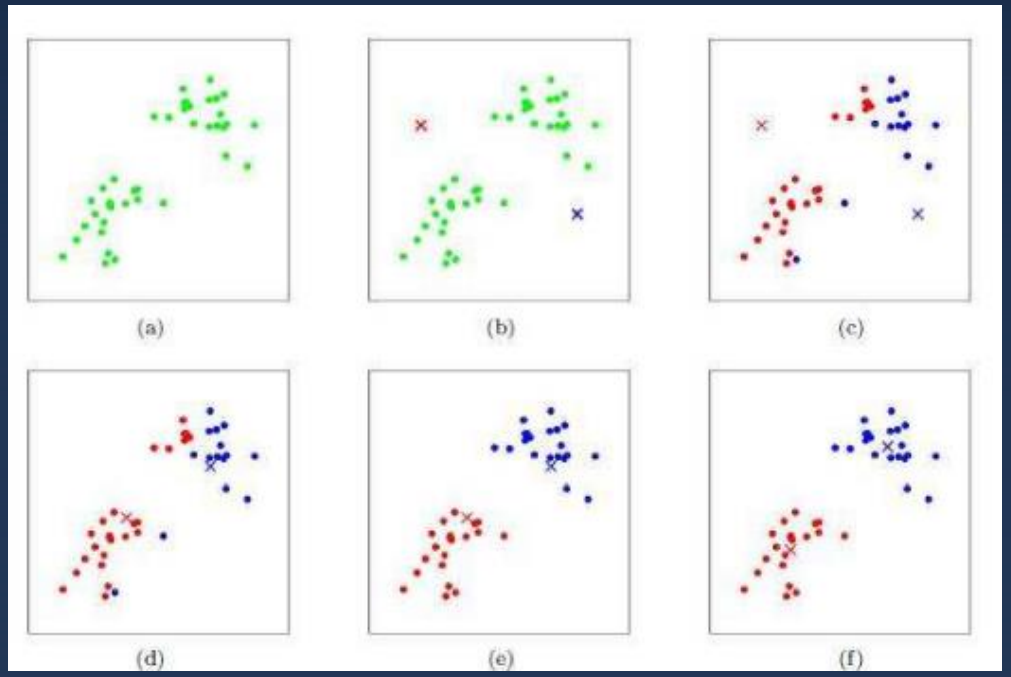
方法： 基于划分聚类有可能会要穷举所有可能的划分。可以采用启发式的方法，逐渐提高簇的质量，逼近局部最优解。这种基于划分的方法非常适合求解球形簇。

特点：

1、发现高维空间中球形的互斥的簇。
2、基于距离。
3、可以使用簇均值或者中心点代表簇。

划分聚类算法：k-means 原理

输入：k（簇的数目）、要进行分类的数据集D
输出：k个簇的集合
过程：
1）从D中以某种规则选择k个样本或者k个在值域范围内的点作为初始簇的中心
2）计算簇中心之外的每个样本和每个簇中心的距离，将样本归属于最近的的簇。
3）计算簇内均值，将均值作为簇的中心。
4）训练2和3，直到簇中心的变化小于一定的阈值或者即便簇中心变动，但是簇内样本不变动。



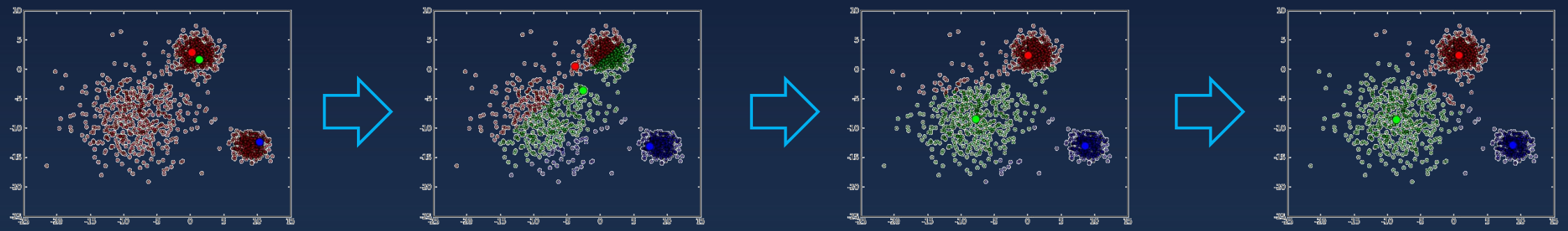
k-means的优缺点：

优点：
1、能够比较快速的收敛。
2、在大数据集上是相对可伸缩和有效的。
3、当无法计算均值时，可以通过定义一些簇中心（比如众数簇心）来改写算法。

缺点：
1、严重依赖于k值的确定。
2、不适合发现非球形簇或者大小差别非常大的簇。
3、对噪声和离群点非常的敏感。

划分聚类算法：k-means 图示

成功案例

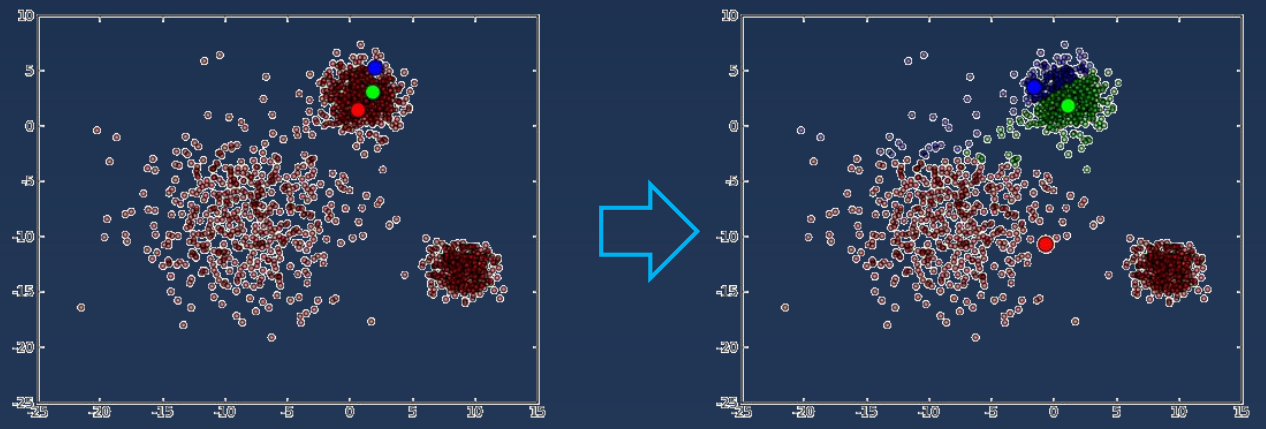


最小化目标函数

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

失败案例



k-means 初始化方法

K值设定：

提供一些基于经验的k值，或者一个k值的范围，通过比较由不同的k得到的聚类结果，来确定最佳的k值。

初始簇心选择：两种方式

1、随机初始簇心：k个簇的簇心的初始化都是随机的，随机选择k个样本作为簇心。

2、有限最大距离簇心：

1）随机选择1个样本作为第1个簇心

2）随机选择m个样本，计算这些样本和当前所有簇心的距离dist

3）在2的计算结果中选择dist最大的样本作为下1个簇心。

4）循环2和3，直到选出k个簇心。

终止规则：两种方式

1、完全终止，即在当前k值设定下，簇完全不发生更新时停止

2、设定一个次数，当簇的更新次数达到此值时即停止

目录

Contents

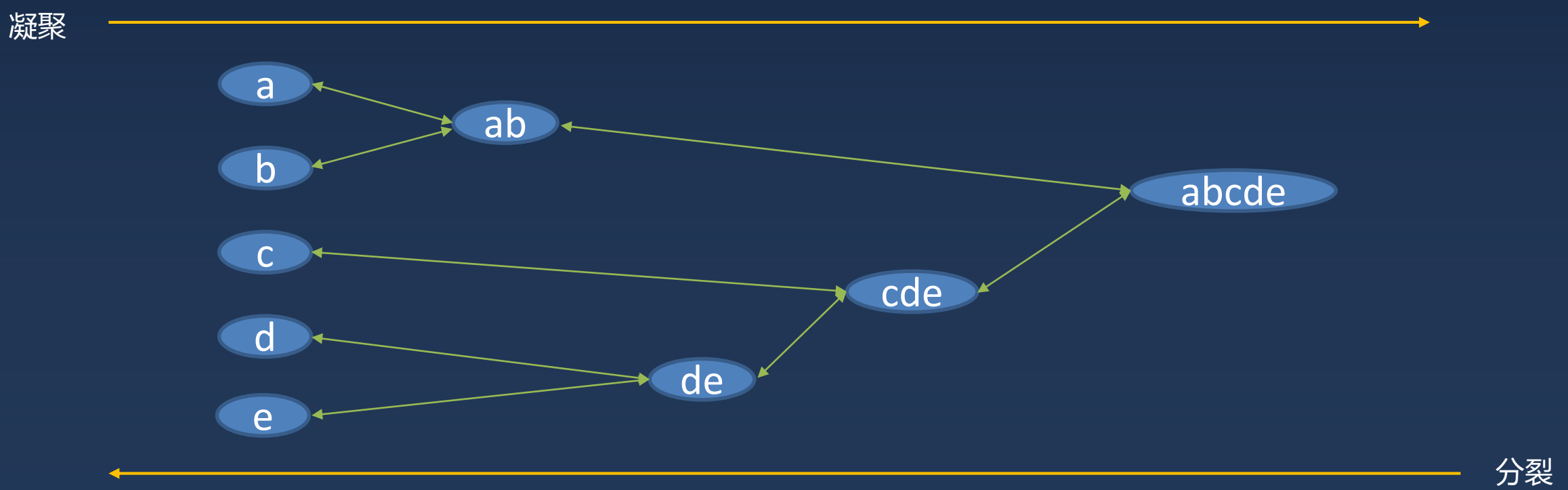
层次方法

在某些情况下，希望可以将数据划分成不同层次或者level上的簇。比如考虑某一种疾病的发病率，可以从个体发病率的角度去考虑、也可以从群体（比如性别）的角度去考虑。如果能够将数据在不同层次上进行聚类，则能发现更多的数据规律。

可以把所有数据看做一个大簇，也可以把每一个样本看做一个小簇。这样就衍生出了基于分裂的层次方法和基于凝聚的层次方法。

基于分裂的层次方法：自顶向下，迭代的进行分裂，从1个大簇形成多个小簇，最底层的簇都足够的凝聚（包含一个对象 or 簇内的样本彼此都充分的相似）

基于凝聚的层次方法：自底向上，迭代的进行合并，从n个小簇形成更大的簇，终止条件可设定，也可以是最终形成一个簇。



样本距离计算方法

样本的凝聚或者分裂的基础是样本间距离的计算，距离最近或者距离在一个阈值范围内则归属于同一个簇，反之则归属于不同的簇。常用的距离计算方法很多，如下列出部分距离的计算公式：

欧式距离：

$$d = \sqrt{\sum_{i=1}^{N}(x_{1i} - x_{2i})^2}$$

标准化欧式距离：

$$d = \sqrt{\sum_{i=1}^{N}\frac{(x_i - y_i)^2}{s_i^2}}$$

S_i是特征i上的标准差

余弦距离：

$$d = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_i^N x_i^2}\sqrt{\sum_i^N y_i^2}}$$

曼哈顿距离：

$$d_{12} = \sum_{k=1}^{n} |x_{1k} - x_{2k}|$$

Thank You.

华为云机器学习服务MLS
www.huaweicloud.com/product/mls.html