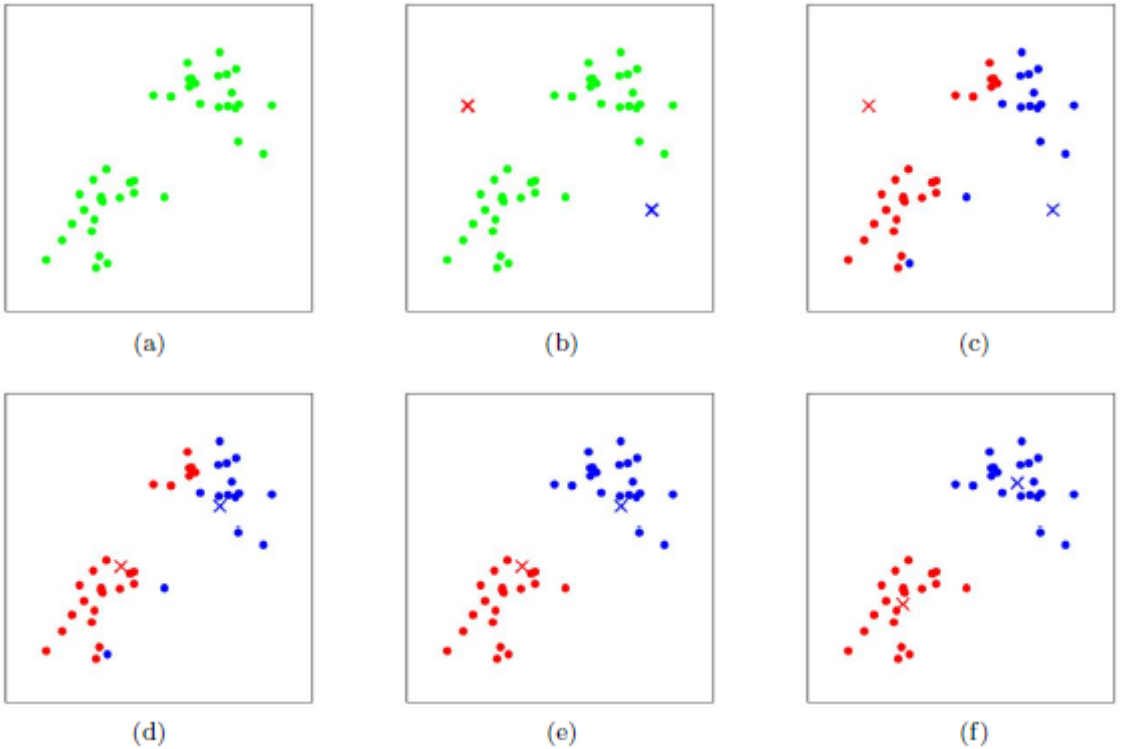
聚类——认识K-means算法

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

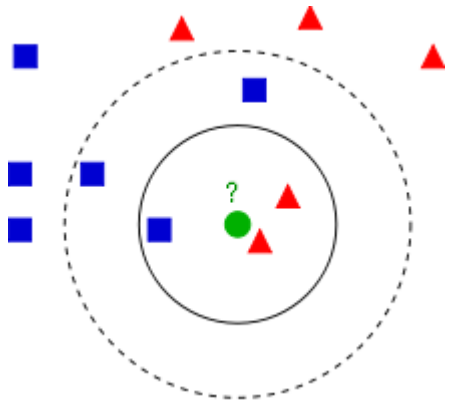
一、聚类与分类

聚类： 无监督学习。聚类是在预先不知道欲划分类的情况下，根据信息相似度原则进行信息聚类的一种方法。 目的是使得属于同类别的对象之间的差别尽可能的小，而不同类别上的对象的差别尽可能的大。

分类： 监督学习，即每个训练样本的数据对象已经有类标识，通过学习可以形成表达数据对象与类标识间对应的知识。 目的是根据样本数据形成的类知识并对源数据进行分类，进而也可以预测未来数据的归类。



聚类分析图（K-means算法）



分类（KNN）

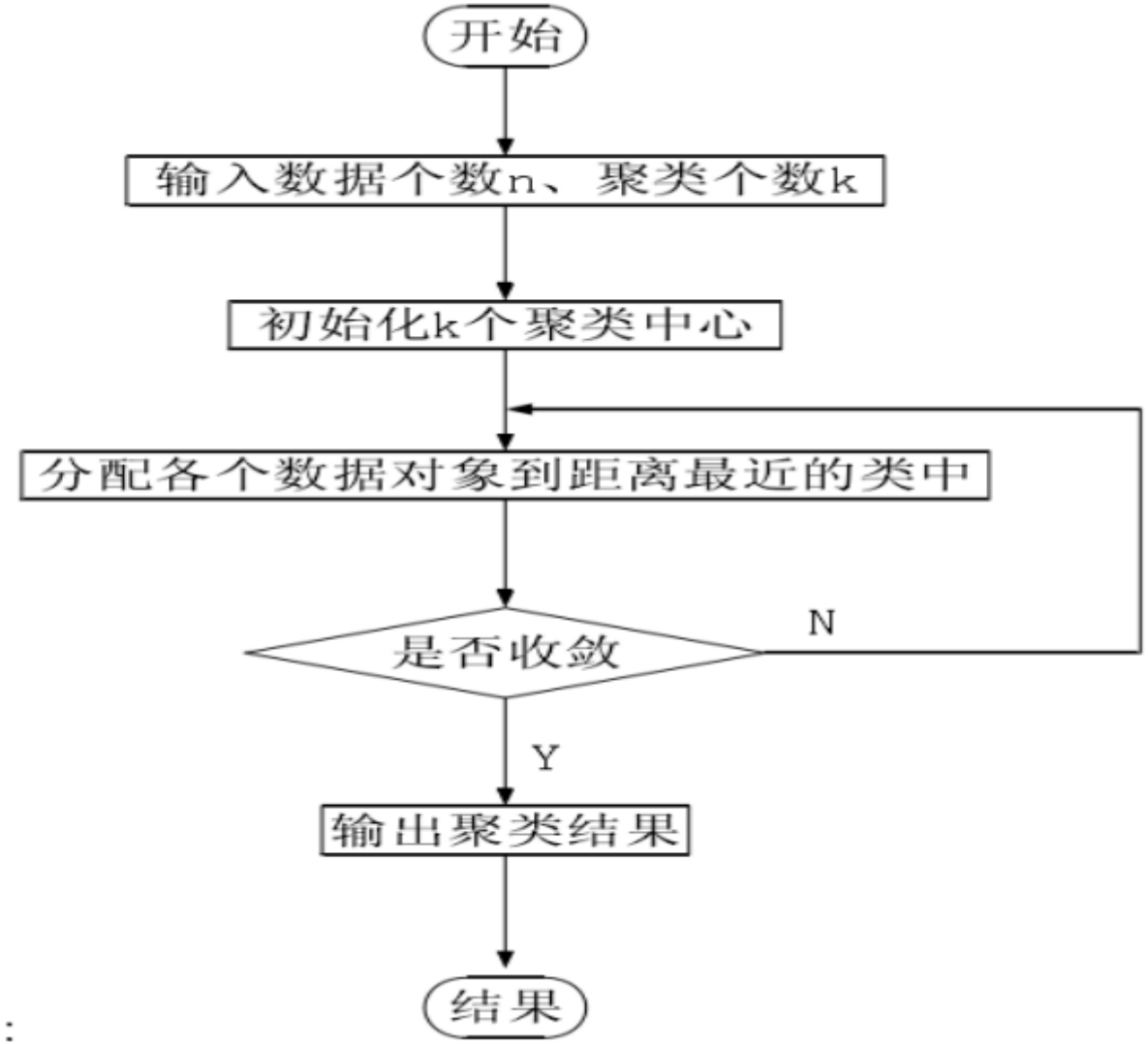
KNN（K-Nearest Neighbor）	K-Means
分类算法	聚类算法
监督学习	非监督学习
数据集是带label的数据，已经是完全正确的数据	数据集是无label的数据，是杂乱无章的，经过聚类后才变得有点顺序，先无序，后有序
没有明显的前期训练过程	有明显的前期训练过程
K的含义：来了一个样本x，要给它分类，即求出它的y，就从数据集中，在x附近找离它最近的K个数据点，这K个数据点，类别c占的个数最多，就把x的label设为c	K的含义：K是人工固定好的数字，假设数据集合可以分为K个簇，由于是依靠人工定好，需要一点先验知识
相似点：都包含这样的过程，给定一个点，在数据集中找离它最近的点。	

二、K-means算法

1.概述

K均值聚类算法是一种经典的划分聚类算法，也是一种迭代的聚类算法，在迭代的过程中不断移动聚类中心，直到聚类准则函数收敛为止。

2.算法实现流程



3.算法步骤

K-means

输入：聚类的数目k和包含n个对象的数据集。

步骤1：从数据中任意选择k个样本数据作为初始聚类中心，表示为 $C=\{c_1,c_2,\cdots,c_k\}$ 。

步骤2：对剩余的每个待聚类数据对象，根据以下公式将样本数据指派到距离最小的类簇中；

$$d(x_j,c_i)=\|x_j-c_i\|_2=(\sum_{l=1}^{p}|x_{jl}-c_{il}|^2)^{\frac{1}{2}}$$

步骤3：根据以下公式更新k个聚类中心的值；

$$c_i=\frac{1}{n}\sum_{x_j\in S_i}x_j$$

步骤4：验证算法是否停止，若满足以下目标函数值最小或保持不变，则迭代结束：否则，执行步骤2。

$$J(X,C)=\sum_{i=1}^{k}\sum_{x_j\in S_i}d(x_j,c_i)$$

推荐：[Clustering - K-means]