一、 DBScan 算法简介

DBSCAN(Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法)是一种很典型的密度聚类算法，它是一种基于高密度连通区域的、基于密度的聚类算法，能够将具有足够高密度的区域划分为簇，并在具有噪声的数据中发现任意形状的簇。

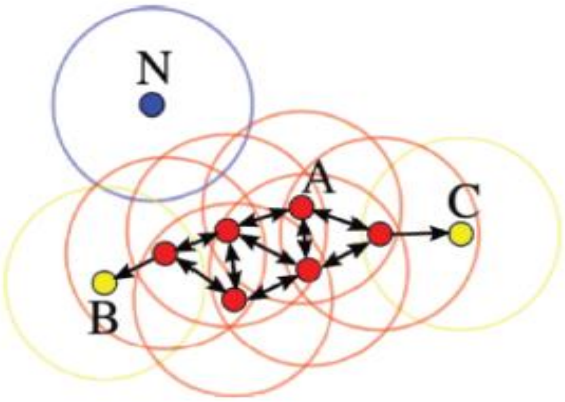
二、 DBScan 算法基本定义

1. ε邻域：给定对象半径ε内的区域称为该对象的ε邻域。
2. 核心对象 :如果给定对象ε邻域内的样本点数大于等于$MinPts$,则称该对象为核心对象。
3. 直接密度可达：给定一个对象集合D，如果p在q的ε邻域内，且 q 是一个核心对象，则我们说对象 p 从对象 q 出发是直接密度可达的。
4. 密度可达：对于样本集合 D，如果存在一个对象链

   $$p_1, p_2, \dots, p_n, p_1 = q, p_n = p$$

   对于$p_i \in D(1 \le i \le n)$，p_{i+1}是从p_i关于ε和 $MinPts$ 直接密度可达，则对象 p 是从对象 q 关于ε和 $MinPts$ 密度可达的。
5. 密度相连：如果存在对象 $o \in D$，使对象 p 和 q 都是从 o 关于ε和 $MinPts$ 密度可达的，那么对象 p 到 q 是关于ε和 $MinPts$ 密度相连的。

密度可达是直接密度可达的传递闭包，并且这种关系是非对称的。只有核心对象之间相互密度可达。然而，密度相连是对称关系。DBSCAN 目的是找到密度相连对象的最大集合。



如图所示，其中红色为核心点，黄色为边界点，蓝色为噪音点。

三、 DBScan 算法流程

DBScan 算法的基础思想是：一个聚类可以由其中的任何核心对象唯一确定。等价可以表述为：任一满足核心对象条件的数据对象 p，数据库 D 中所有从 p 密度可达的数据对象 o 所组成的集合构成了一个完整的聚类 C，且 p 属于 C。因此我们可以得到该算法的流程如下：

扫描整个数据集，找到任意一个核心点，对该核心点进行扩充。扩充的方法是寻找从该核心点出发的所有密度相连的数据点（注意是密度相连）。遍历该核心点的ε邻域内的所有核心点（因为边界点是无法扩充的），寻找与这些数据点密度相连的点，直到没有可以扩充的数据点为止。最后聚类成的簇的边界节点都是非核心数据点。之后就是重新扫描数据集（不包括之前寻找到的簇中的任何数据点），寻找没有被聚类的核心点，再重复上面的步骤，对该核心点进行扩充直到数据集中没有新的核心点为止。数据集中没有包含在任何簇中的数据点就构成异常点。

四、 轮廓系数

轮廓系数是一个进行聚类算法模型评估的重要指标，它的描述如下：

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}},s(i)=\begin{cases}1-\frac{a(i)}{b(i)},a(i)<b(i)\\0,a(i)=b(i)\\\frac{b(i)}{a(i)}-1,a(i)>b(i)\end{cases}$$

- 计算样本 i 到同簇其他样本的平均距离$a(i)$。$a(i)$越小，说明样本 i 越应该被聚类到该簇。将$a(i)$称为样本 i 的簇内不相似度。
- 计算样本 i 到其他某簇的平均距离$b(ij)$ 称为样本 i 与某簇的不相似度。定义为样本 i 的簇间不相似度，$b(i)=min\{b(i1),b(i2),\dots,b(ik)\}$
- $s(i)$接近 1，则说明样本 i 聚类合理
- $s(i)$接近-1，说明样本 i 应该分类到其他的簇
- $s(i)$近似为哦，说明样本 i 在两个簇的边界上。

五、 DBScan 算法的优缺点

优点

- 聚类速度快且能够有效处理噪声点和发现任意形状的空间聚类；
- 不需要输入要划分的聚类个数；
- 聚类簇的形状没有偏倚；
- 可以在需要时输入过滤噪声的参数。

缺点

- 当数据量增大时，要求较大的内存支持 I/O 消耗也很大；
- 当空间聚类的密度不均匀、聚类间距差相差很大时，参数 MinPts 和 Eps 选取困难，聚类质量较差。
- 算法聚类效果依赖与距离公式选取，实际应用中常用欧式距离，对于高维数据，存在“维数灾难”。