### 4.1.1 **DBSCAN**聚类算法

DBSCAN（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法）一种很典型的密度聚类算法，和K-Means，BIRCH这些一般只适用于凸样本集的聚类相比，DBSCAN既可以适用于凸样本集，也可以适用于非凸样本集。

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，即要求在聚类空间中的一定区域内所包含对象(点或其他空间对象)的数目不小于某一给定阈值。过滤低密度区域，发现稠密度样本点。同一类别的样本，他们之间是紧密相连的，即在该类别任意样本周围不远处一定有同类别的样本存在。下面介绍DBSCAN算法的基本定义以及思路。

假设当前有假设样本集是

1. ϵ-邻域：对于，其ϵ-邻域包含样本集D中与的距离不大于ϵ的子样本集。
2. 核心对象：对于任一样本，如果其ϵ-邻域对应的子样本集个数至少包含MinPts个样本，则是核心对象。
3. 密度直达：如果位于的ϵ-邻域中，且是核心对象，则称由密度直达。
4. 密度可达：对于和,如果存在样本序列满足，, 且，由密度直达，则称由密度可达。也就是说，密度可达满足传递性。此时序列中的传递样本均为核心对象，因为只有核心对象才能使其他样本密度直达。
5. 密度相连：对于和，如果存在核心对象样本，使和均由密度可达，则称和密度相连。

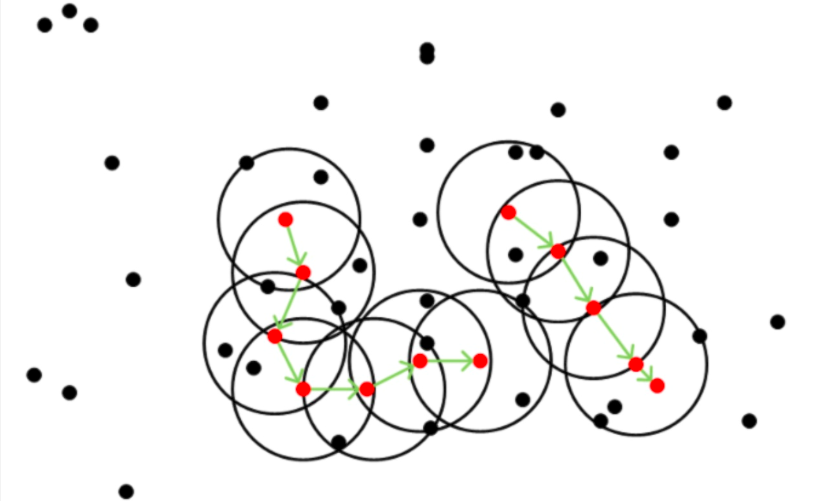


图4.1 DBSCAN图示

 从上图4.1可以很容易看出理解上述定义，图中MinPts=5，红色的点都是核心对象，因为其ϵ-邻域至少有5个样本。黑色的样本是非核心对象。所有核心对象密度直达的样本在以红色核心对象为中心的超球体内，如果不在超球体内，则不能密度直达。图中用绿色箭头连起来的核心对象组成了密度可达的样本序列。在这些密度可达的样本序列的ϵ-邻域内所有的样本相互都是密度相连的。

由密度可达关系导出的最大密度相连的样本集合，即为我们最终聚类的一个类别，或者说一个簇。这个DBSCAN的簇里面可以有一个或者多个核心对象。如果只有一个核心对象，则簇里其他的非核心对象样本都在这个核心对象的ϵ-邻域里;如果有多个核心对象，则簇里的任意一个核心对象的ϵ-邻域中一定有一个其他的核心对象，否则这两个核心对象无法密度可达。这些核心对象的ϵϵ-邻域里所有的样本的集合组成的一个DBSCAN聚类簇。

下面是具体的实验步骤：

1、DBSCAN算法从一个未被访问的任意的数据点开始。这个点的邻域是用距离epsilon来定义（即该点ε距离范围内的所有点都是邻域点）。

2、如果在该邻域内有足够数量的点（根据minPoints的值），则聚类过程开始，并且当前数据点成为新簇中的第一个点。否则，该点将被标记为噪声（稍后，这个噪声点可能成为聚类中的一部分）。在这两种情况下，该点都会被标记为“已访问”。

3、对于新簇中的第一个点，它的ε距离邻域内的点也会成为同簇的一部分。这个过程使ε邻域内的所有点都属于同一个簇，然后对才添加到簇中的所有新点重复上述过程。

4、重复步骤2和3两个过程直到确定了聚类中的所有点才停止，即访问和标记了聚类的ε邻域内的所有点。

5、一旦我们完成了当前的聚类，就检索和处理新的未访问的点，就能进一步发现新的簇或者是噪声。重复上述过程，直到所有点被标记为已访问才停止。由于所有点已经被访问完毕，每个点都被标记为属于一个簇或是噪声。

在此算法中有两个关键点，一个是ϵ 某一样本的邻域距离阈值（即领域半径）一个是MinPts 某一样本的距离为 ϵ 的邻域中样本个数的阈值（即最少点个数）。将其与数据压缩相结合，即ϵ 为列向量之间的欧几里得距离。原测试集中如果有大于等于MinPts个列向量与当前列向量之间的欧几里得距离小于ϵ，就可以确定当前列向量为一个中心点。根据电路的大小，需要选取的基向量数目为k，那么可以设计合理的ϵ 和MinPts确定聚类中心即可。