机器学习算法系列（11）：聚类（3）— DBSCAN

四、DBSCAN算法

4.1 密度聚类方法

密度聚类方法的指导思想是，只要样本点的密度大于某阈值，则将该样本添加到最近的簇中。这类算法能克服基于距离的算法只能发现“类圆”（凸）的聚类的缺点，可发现任意形状的聚类，且对噪声数据不敏感。但计算密度单元的计算复杂度大，需要建立空间索引来降低计算量。
其代表算法为DBSCAN算法和密度最大值算法。

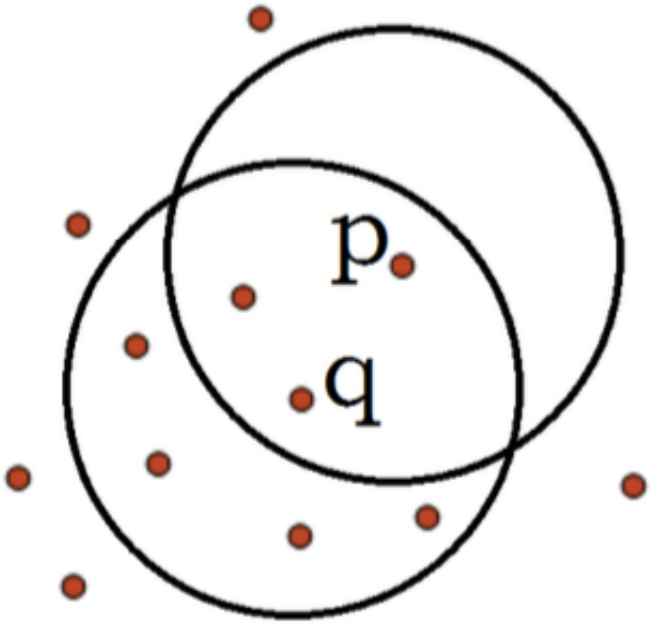
4.2 DBSCAN算法原理

DBCSAN（Density-Based Spatial Clustering of Applications with Noise）是一个比较有代表性的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在有“噪声”的数据中发现任意形状的聚类。

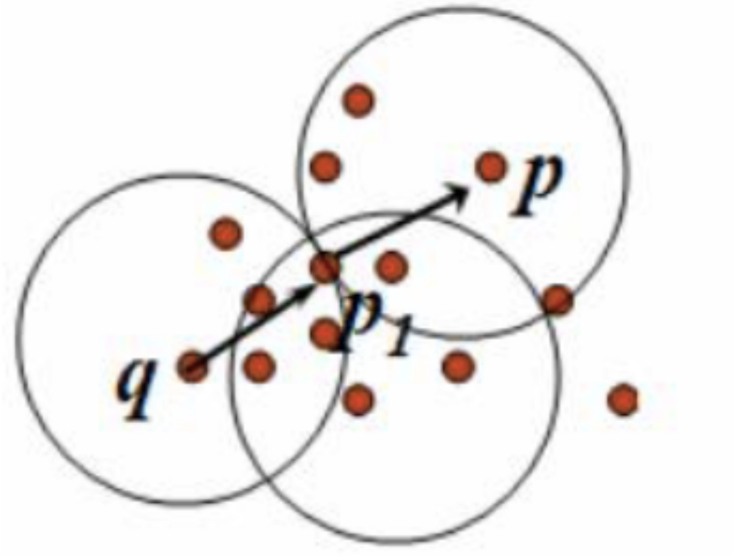
4.3 若干概念

- 对象的$\varepsilon-$领域：给定对象在半径ε内的区域
- 核心对象：对于给定的数目m，如果一个对象的$\varepsilon-$领域至少包含m个对象，则称该对象为核心对象。

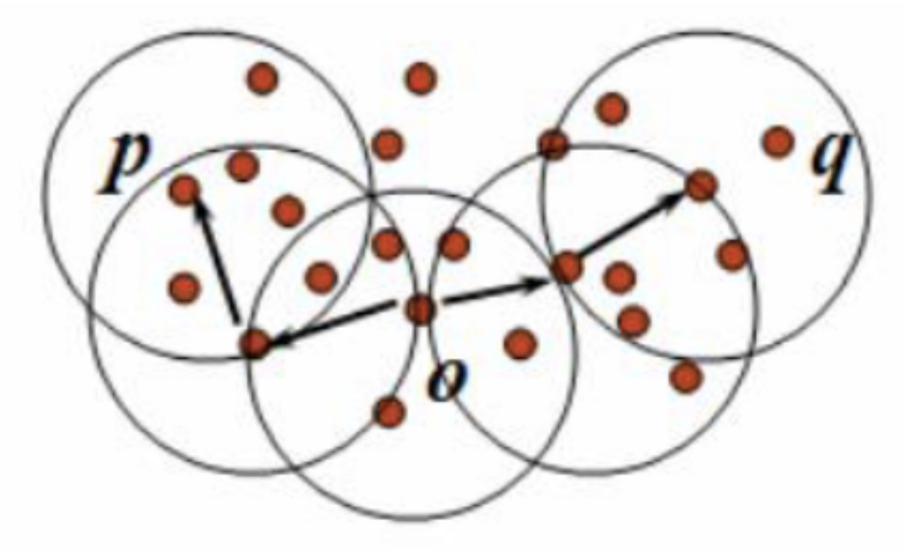
- 直接密度可达：给定一个对象集合D，如果p是在q的$\varepsilon-$领域内，而q是一个核心对象，我们说对象p从对象q出发时直接密度可达的。
  如图$\varepsilon=1, m=5$，q是一个核心对象，从对象q出发到对象p是直接密度可达的。



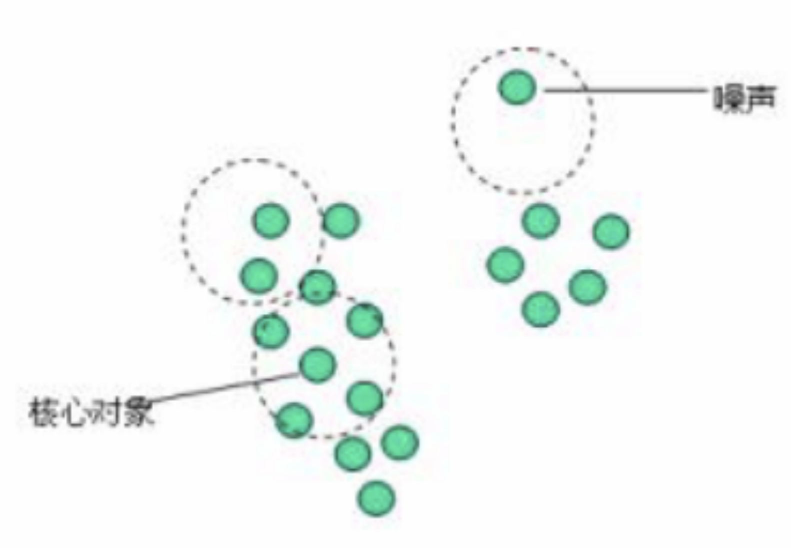
- 密度可达：如果存在一个对象链$p_1 p_2 \cdots p_n$，$p_1 = q, p_n = p$，对$p_i \in D, (1 \leq i \leq n)$，$p_{i+1}$是从$p_i$关于$\varepsilon$和$m$直接密度可达的，则对象$p$是从对象$q$和$m$密度可达的。



- 密度相连：如果对象集合D中存在一个对象O，使得对p和q是从O关于ε和m密度可达的，那么对象p和q是关于ε和m密度相连的。

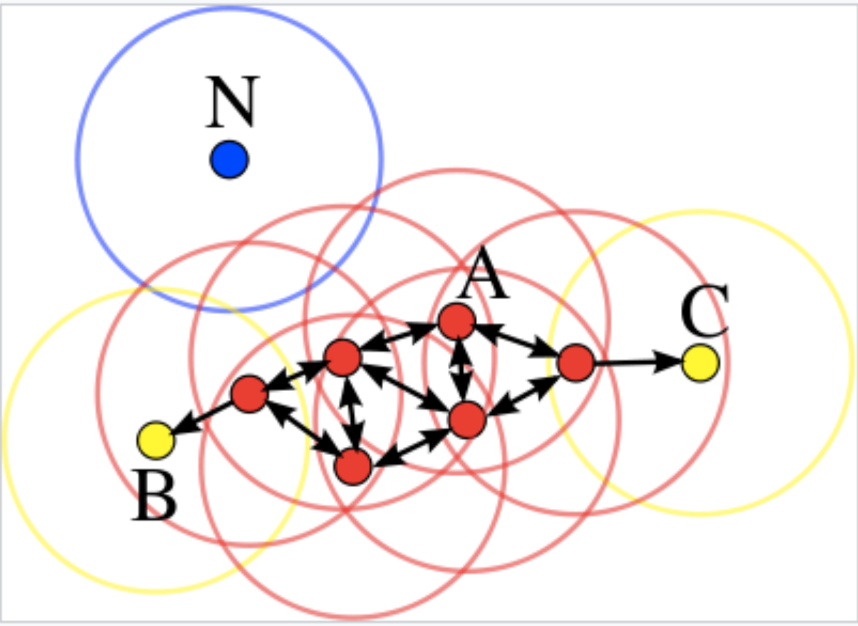


- 簇：一个基于密度的簇是最大的密度相连对象的集合。
- 噪声：不包含在任何簇中的对象称为噪声。



4.4 算法步骤

下面这张图来自WIKI，图上有若干个点，其中标出了A、B、C、N这四个点，据此来说明这个算法的步骤：



In this diagram, $\mathrm{minPts} = 4$. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor density-reachable.

- 1、首先随机选择A点为算法实施的切入点，我们将ε设置为图中圆的半径，对象个数m（$minPts$）设定为4。这里我们看到，A点的$\varepsilon-$领域包含4个对象（自己也包含在内），大于等于$m(minPts)$，则创建A作为核心对象的新簇，簇内其他点都（暂时）标记为边缘点。
- 2、然后在标记的边缘点中选取一个重复上一步，寻找并合并核心对象直接密度可达的对象。对暂时标记为边缘点反复递归上述算法，直至没有新的点可以更新簇时，算法结束。这样就形成了一个以A为起始的一个聚类，为图中红色的中心点和黄色的边缘点
- 3、如果还有Points未处理，再次新产生一个类别来重新启动这个算法过程。遍历所有数据，如果有点既不是边缘点也不是中心点，将其标记为噪音。

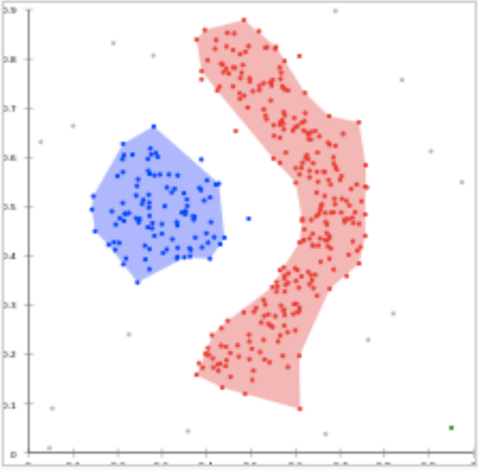
从上述算法可知：

- 每个簇至少包含一个核心对象；
- 非核心对象可以是簇的一部分，构成了簇的边缘（edge）；
- 包含过少对象的簇被认为是噪声；

4.5 总结

- 优点
    - 无需确定聚类个数：DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-means.

  - 可以发现任意形状的聚类：DBSCAN can find arbitrarily shaped clusters. It can even find a cluster completely surrounded by (but not connected to) a different cluster. Due to the MinPts parameter, the so-called single-link effect (different clusters being connected by a thin line of points) is reduced.
  - 对噪声具有鲁棒性，可有效处理噪声：DBSCAN has a notion of noise, and is robust to outliers.
  - 只需两个参数，对数据输入顺序不敏感：DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database. (However, points sitting on the edge of two different clusters might swap cluster membership if the ordering of the points is changed, and the cluster assignment is unique only up to isomorphism.)
  - 加快区查询：DBSCAN is designed for use with databases that can accelerate region queries, e.g. using an R* tree.
  - 参数可由领域专家设置：The parameters minPts and ε can be set by a domain expert, if the data is well understood.



DBSCAN can find non–linearly separable clusters. This dataset cannot be adequately clustered with k–means or Gaussian Mixture EM clustering.

- 缺点

  - 边界点不完全确定性：DBSCAN is not entirely deterministic: border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data is processed. Fortunately, this situation does not arise often, and has little impact on the clustering result[citation needed]: both on core points and noise points, DBSCAN is deterministic. DBSCAN*[4] is a variation that treats border points as noise, and this way achieves a fully deterministic result as well as a more consistent statistical interpretation of density-connected components.

- 维数灾导致欧几里得距离度量失效：The quality of DBSCAN depends on the distance measure used in the function regionQuery(P,ε). The most common distance metric used is Euclidean distance. Especially for high-dimensional data, this metric can be rendered almost useless due to the so-called "Curse of dimensionality", making it difficult to find an appropriate value for ε. This effect, however, is also present in any other algorithm based on Euclidean distance.
- 不能处理密度差异过大（密度不均匀）的聚类（会导致参数无法适用于所有聚类）：DBSCAN cannot cluster data sets well with large differences in densities, since the minPts-ε combination cannot then be chosen appropriately for all clusters.
- 参数选择在数据与规模不能很好理解的情况下，很难选择，若选取不当，聚类质量下降： If the data and scale are not well understood, choosing a meaningful distance threshold ε can be difficult.