sklearn中聚类算法测试实验报告

1. 实验任务

测试sklearn中K-Means, Affinity Propagation, Mean-Shift, Spectral,Ward hierarchical, Agglomerative,DBSCAN, Gaussian Mixture八种聚类算法在tweets数据集上的聚类效果。并使用NMI(Normalized Mutual Information)作为评价指标。

1. 算法简介

**K-Means：**

①需要选取k个初始质心作为初始cluster 对每个样本点；

②计算得到距其最近的质心，将其类别标为该质心所对应的cluster； 重新计算k个cluster对应的质心；

③重复②直到质心不再发生变化。

**Affinity Propagation：**

  AP聚类算法是基于数据点间的"信息传递"的一种聚类算法。与k-均值算法不同，AP算法不需要在运行算法之前确定聚类的个数。AP算法寻找的聚类中心点是数据集合中实际存在的点，作为每类的代表。

**Mean-Shift：**

Mean-Shift聚类法可以自动确定k的个数，下面简要介绍一下其算法流程：

①随机确定样本空间内一个半径确定的高维球及其球心；

②求该高维球内质心，并将高维球的球心移动至该质心处；

③重复②，直到高维球内的密度随着继续的球心滑动变化低于设定的阈值，算法结束。

**Spectral：**

谱聚类主要思想是把所有的数据看做空间中的点，这些点之间可以用边连接起来。距离较远的两个点之间的边权重值较低，而距离较近的两个点之间的边权重值较高，通过对所有数据点组成的图进行切图，让切图后不同的子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到聚类的目的。

**Ward hierarchical：**

层次聚类试图在不同的“层次”上对样本数据集进行划分，一层一层地进行聚类。自底向上的凝聚方法（agglomerative hierarchical clustering）是先将所有样本的每个点都看成一个簇，然后找出距离最小的两个cluster进行合并，不断重复到预期cluster或者其他终止条件。其中，ward是一种链接方式。

**Agglomerative**：

使用自底向上的凝聚方法的层次聚类，除ward链接方法外，还有complete和average。

**DBSCAN：**

DBSCAN是一种基于密度的聚类算法，这类密度聚类算法一般假定类别可以通过样本分布的紧密程度决定。同一类别的样本，他们之间的紧密相连的，也就是说，在该类别任意样本周围不远处一定有同类别的样本存在。　　　　通过将紧密相连的样本划为一类，这样就得到了一个聚类类别。通过将所有各组紧密相连的样本划为各个不同的类别，则我们就得到了最终的所有聚类类别结果。

**Gaussian Mixture：**

Gaussian Mixture是用高斯概率密度函数精确地量化事物，将一个事物分解为若干的基于高斯概率密度函数形成的模型。无论观测数据集如何分布以及呈现何种规律，都可以通过多个单一高斯模型的混合进行拟合。

1. 实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 方法 | NMI | 簇数 | 是否预先给定簇数 |
| K-Means | 0.78 | 89 | 是 |
| Affinity propagation | 0.72 | 354 | 否 |
| Mean-Shift | 0.74 | 766 | 否 |
| Spectral | 0.68 | 89 | 是 |
| Agglomerative（Ward） | 0.78 | 89 | 是 |
| Agglomerative（Complete） | 0.74 | 89 | 是 |
| Agglomerative（average） | 0.90 | 89 | 是 |
| DBSCAN | 0.63 | 95 | 否 |
| Gaussian mixture | 0.78 | 89 | 是 |

实验心得：

对于聚类来讲，参数的选择对于聚类的效果至关重要。