Data Mining Homework3 总结报告

1、作业要求

在 tweets 数据集上利用sklearn进行各种聚类算法的实验，然后用 NMI(Normalized Mutual Information)作为评价指标对聚类结果进行评价。

2、数据处理

读取tweets文件，解析文件json格式，并用sklearn的TfidfVectorizer、CountVectorizer进行数据聚类预处理。

3、测试聚类算法

（1）K-Means

K-means算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

使用sklearn的KMeans进行聚类，参数K=89，结果如下：

|  |  |
| --- | --- |
| 第 0 次 Kmeans NML = 0.791697 |  |
| 第 1 次 Kmeans NML = 0.779661 |  |
| 第 2 次 Kmeans NML = 0.809423 | 最高 |
| 第 3 次 Kmeans NML = 0.774398 | 最低 |
| 第 4 次 Kmeans NML = 0.782131 |  |
| 第 5 次 Kmeans NML = 0.792177 |  |
| 第 6 次 Kmeans NML = 0.794749 |  |
| 第 7 次 Kmeans NML = 0.790386 |  |
| 第 8 次 Kmeans NML = 0.791627 |  |
| 第 9 次 Kmeans NML = 0.782180 |  |

（2）Affinity Propagation

Affinity Propagation聚类算法简称AP，算法的基本思想是将全部样本看作网络的节点，然后通过网络中各条边的消息传递计算出各样本的聚类中心。

使用sklearn的AffinityPropagation聚类，结果如下：

Affinity Propagation NML = 0.7834。

（3）Mean-Shift

Mean Shift算法是一种无参密度估计算法或称核密度估计算法，Mean shift是一个向量，它的方向指向当前点上概率密度梯度的方向。所谓的核密度评估算法，指的是根据数据概率密度不断移动其均值质心（也就是算法的名称Mean Shift的含义）直到满足一定条件。

使用sklearn的MeanShift聚类，参数bandwidth=9，结果如下：

MeanShift NML = 0.6992。

（4）Spectral Clustering

谱聚类是从图论中演化出来的算法，后来在聚类中得到了广泛的应用。它的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。

使用sklearn的SpectralClustering聚类，结果如下：

Spectral Clustering NML = 0.6703。

（5）Ward Hierarchical Clustering

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。就划分策略可分为自底向上的凝聚方法（agglomerative hierarchical clustering），自上向下的分裂方法（divisive hierarchical clustering）。

使用sklearn的AgglomerativeClustering聚类，参数K=89，结果如下：

Ward Hierarchical Clustering NML = 0.7800。

（6）Agglomerative Clustering

Agglomerative Clutsering 是一种自底而上的层次聚类方法，它能够根据指定的相似度或距离定义计算出类之间的距离。依据对相似度（距离）的不同定义，将Agglomerative Clustering的聚类方法分为三种：Single-linkage,Complete-linkage和Group average。Single-linkage:要比较的距离为元素对之间的最小距离；Complete-linkage:要比较的距离为元素对之间的最大距离；Group average：要比较的距离为类之间的平均距离（平均距离的定义与计算：假设有A，B两个类，A中有n个元素，B中有m个元素。在A与B中各取一个元素，可得到他们之间的距离。将nm个这样的距离相加，得到距离和。最后距离和除以nm得到A，B两个类的平均距离）。

使用sklearn的AgglomerativeClustering聚类，参数linkage='complete'、K=89，结果如下：

Agglomerative Clustering NML = 0.7424。

（7）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。

使用sklearn的DBSCAN聚类，参数eps=0.1、 min\_samples=1，结果如下：

DBSCAN NML = 0.6985。

（8）Gaussian Mixtures

Gaussian Mixtures（混合高斯模型）就是指对样本的概率密度分布进行估计，而估计的模型是几个高斯模型加权之和（具体是几个要在模型训练前建立好）。每个高斯模型就代表了一个类（一个Cluster）。对样本中的数据分别在几个高斯模型上投影，就会分别得到在各个类上的概率。然后我们可以选取概率最大的类所为判决结果。

使用sklearn的DBSCAN聚类，参数n\_components=89、 covariance\_type='diag'，进行10次聚类分析结果如下：

|  |  |
| --- | --- |
| 第 0 次 Gaussian Mixtures NML = 0.7681 | 最低 |
| 第 1 次 Gaussian Mixtures NML = 0.7873 |  |
| 第 2 次 Gaussian Mixtures NML = 0.7845 |  |
| 第 3 次 Gaussian Mixtures NML = 0.7969 |  |
| 第 4 次 Gaussian Mixtures NML = 0.7720 |  |
| 第 5 次 Gaussian Mixtures NML = 0.7815 |  |
| 第 6 次 Gaussian Mixtures NML = 0.7867 |  |
| 第 7 次 Gaussian Mixtures NML = 0.7832 |  |
| 第 8 次 Gaussian Mixtures NML = 0.7979 | 最高 |
| 第 9 次 Gaussian Mixtures NML = 0.7978 |  |

吉晓辉

2019年1月4日