Programming:

4.1 KMeans，hierarchical clustering，spectral clustering 三种算法的时间复杂度分析。

设数据集的样本个数为 n，维数为 d，聚类个数为 c。

KMeans：

1. 随机选择 c 个中心点
2. 遍历所有数据，将每个数据划分到最近的中心点中
3. 重新计算每个聚类的平均值，并作为新的中心点
4. 重复 2-3 直至收敛

由上述流程可知，时间复杂度为 O(ldcn)，其中 l 为迭代次数，d 为数据维度，c 为类别，n 为样本个数，所以为线性时间复杂度 O(n). 实验验证结果如图 1 所示，具体数据如下表所示。

样本数 n	10	100	1000	10000	60000
时间 s	0.051	0.089	0.794	8.621	80.233

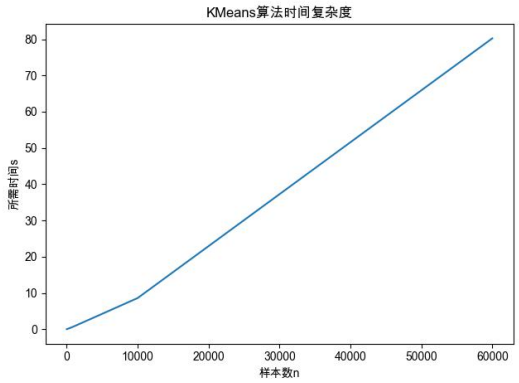

图 1：KMeans 算法时间复杂度

Hierarchical clustering：

1.初始 n 个样本分成 n 类，在所有集合中找到一对满足相似性度量的集合 D_i, D_k

2.将 D_i 并入 D_k

3.当只剩 c 个类别时停止

由上述流程可知，计算相似性度量需要 O(n*n)，共需迭代(n-c)次，所以算法时间复杂度为 $O(n^3)$。60000 会超出计算能力，所以数据如下表所示，实验验证结果如图 2 所示。

样本数 n	10	100	1000	5000	10000
时间 s	0.020	0.055	0.478	11.667	49.449

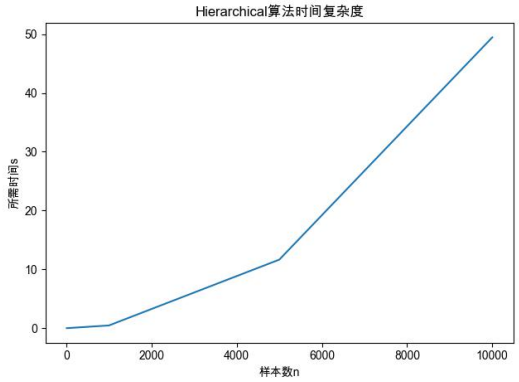

图 2：Hierarchical 算法时间复杂度

spectral clustering:

求矩阵特征值为主要计算，时间复杂度为 $O(n^3)$。具体数据如下表所示，实验验证结果如图 3 所示，算法的时间复杂度比 Hierarchical 算法还要高。

样本数 n	10	50	100	500	1000
时间 s	0.045	0.039	0.131	11.958	63.023

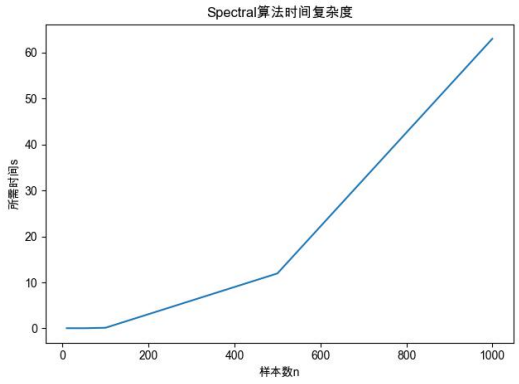

图 3：Spectral 算法时间复杂度

4.2 n_clusters = 10.

(1). 在 KMeans 算法中，初始分割会影响聚类的结果，可能会收敛到局部最优解。设置样本数 n=10000， init = ‘random’，随机选择样本点作为初始中心点。得到结果如下表所示，可以看到初始分割影响收敛时间，最后的 Je 和 NMI 基本相同。但是 Je 和 NMI 保持了对应关系，Je 大 NMI 也大。

	1	2	3	4	5
时间 s	9.281	8.011	8.008	8.633	10.248
Je $*10^5$	3.88135	3.88131	3.88134	3.88141	3.88135
NMI	0.4769	0.4787	0.4775	0.4767	0.4791

解决初始分割的问题，可以设置 init=‘k-means++’，将中心点初始化为彼此远离的状态，可以得到比随机初始化更好的结果。

(2)在 Hierarchical 聚类算法中，设置样本数 n=1000，不同的 linkage 方法会有不同的 NMI 结果。具体数据如下，可以得到在 NMI 定义下，ward 方法最优。

Linkage	ward	complete	average	single
NMI	0.543	0.306	0.277	0.018

(3)在 Spectral 聚类算法中，设置样本数为 n=100，affinity = ‘nearest_neighbors’，n_neighbors = [5,10,20,30]，不同的相似性图和对应的参数会有不同的结果，设置具体数据如下。可以看到近邻法比 rbf 核要好一些，NMI 值更高。而不同的邻居数对 NMI 影响不大。

affinity	rbf	5-nn	10-nn	20-nn	30-nn
NMI	0.534	0.664	0.675	0.675	0.657

4.3 在不知道真实类别数的情况下：

(1) 在 KMeans 算法中，可以在允许的范围内逐个尝试，做一条 Je-c 的函数曲线，其拐点对应的类别数就比较接近于最优类别数。结果如图 4 所示，但是并没有明显的拐点，这个方法不成立。需要运用对数据集已知的知识来分析，确定分类数。例如 MNIST 手写数字集，那么则为 0-9 十个数字，分成 10 类。

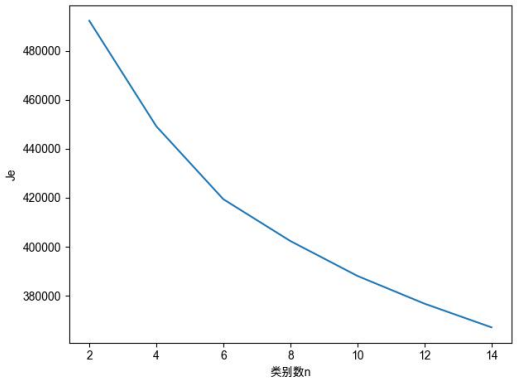

图 4：Je-n 曲线

(2) 在 Hierarchical 算法中，可以画出聚合图，把类间相似度的计算值也在图上加以标明，那么就可以帮助我们确定合理的聚类类数。Samples = 10000 时，画出最后五层聚合图如图 5 所示，如果类别数太少，distance 就会过大，所以可以根据 distance 分类情况确定合适的分类数。

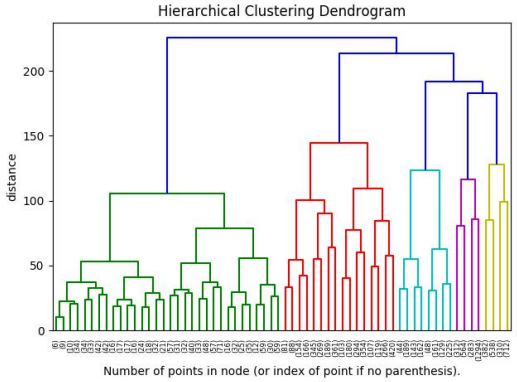

图 5：分层聚合图

4.4 以上几种聚类方法各有优劣，聚类方法的选择要根据样本的分布特性和数量综合考虑。如果样本点成团状分布或者样本数很大时，则用 KMeans 算法能取得较好效果，且速度快，在线性时间复杂度内就能解决；而分层聚类和谱聚类在样本数很大时由于时间复杂度过大而无法使用。如果样本数据较小，且样本点分布不是团状时，则用谱分解更好，因为其相似度图和相似性度量的类型和参数可调，更灵活，也更容易找到与样本点分布对应的聚类结果。