data_mining_hw3

任务二:地理位置数据聚类分析

评价指标

1. 纯度，越接近1越好

$$P=(\Omega,\mathbb{C})=\frac{1}{N}\sum_{k}\max_{j}|\omega_k\cap c_j|$$

2. 兰德指数。TP表示两个相同样本在同一聚类里。TN表示不同样本在两个不同聚类里。

$$RI=\frac{TP+TN}{TP+FP+FN+TN}$$

预处理

聚类的输入数据中无空表项，并且经纬度均符合经纬度的定义限制，无序数据清洗。

聚类算法选择及实验结果

选取了KMeans、DBSCAN、Spectral三种聚类算法。其中自己实现了KMeans的全过程。

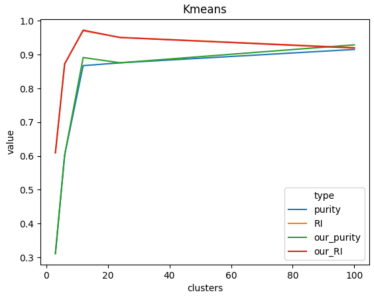
在上述评价指标下得到实验结果如下（其中DBSCAN、Spectral均采用sklearn默认参数、KMeans采取和输入数据类别相同的12为聚类数量，迭代次数为100）：

算法	purity	rand_index
KMeans(Ours)	0.8723459684772698	0.9654111891105196
KMeans(sklearn)	0.8713038947505536	0.9639436177837565
DBSCAN	0.7555034518692197	0.9285248880290016
Spectral	0.7607138205028006	0.9122340012121778

从表格中可以看到，在这两种评价指标和当前的超参数选择上DBSCAN和Spectral接近都差于KMeans。

KMeans实现对比

选取了聚类数量分别为[3,6,12,24,100]对两种KMeans上进行实验得到如下图所示的结果。可以看到在这两种评价指标上和不同K值下，两种算法的结果均接近，甚至在部分情况下我们的实现效果好于sklearns



聚类结果可视化

预处理

由于同一类别间点的经纬度差距相较于不同类别间的经纬度差值实在过小，为了使可视化结果更直观。采用了sklearn.preprocessing里的KBinsDiscretizer将这些点离散化映射到100x100的空间中。

可视化结果

不同K值下聚类结果如下（1,6,12,24,100。仅列出我们的实现的可视化结果），可以看到聚类效果较为明显：