第5次作业\_非监督学习

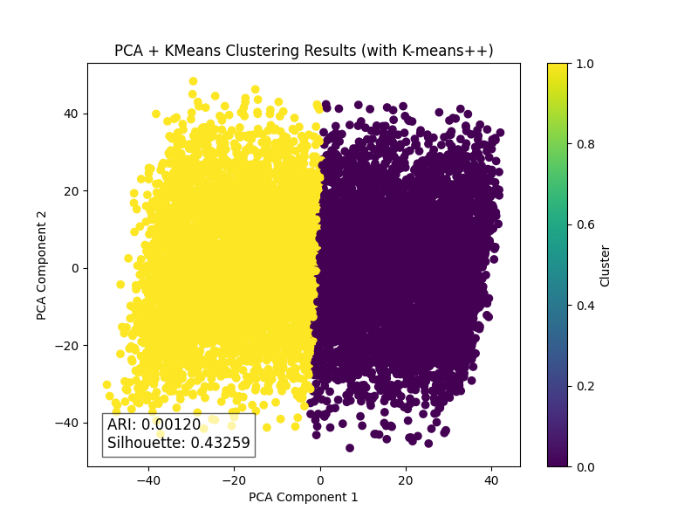
电 25 吴晨聪 2022010311

# （1） 基于数据集提供的系统运行状态，尝试聚类或降维算法。

本次作业的基本思路是利用PCA和t-SNE两种降维算法，再使用 KMeans 算法尝试对系统运行状态聚成 2 类（假设运行状态分为稳定和不稳定）。然后通过观察调整兰德指数 (ARI)和轮廓系数(Silhouette Score)来评价算法的效果。

1. PCA + KMeans

PCA算法注重数据的全局方差，适合线性分布的特征，把参数n\_components设置为2，保留全局方差最大的信息，用于聚类和可视化。下图为聚类结果及与真实SSSA标签的对比图。

一張含有 螢幕擷取畫面, 鮮豔, 文字 的圖片

自動產生的描述

图1 可视化PCA结果与K-means聚类标签以及可视化PCA结果与真实SSSA标签

1. t-SNE + KMeans

t-SNE算法强调局部结构，适合发现非线性特征关系，把参数perplexity设置为50：平衡局部和全局信息，适合电力系统这类数据较多的情况，同时learning\_rate设置为200，较高学习率能加快收敛，把数据降维到 2 维，保留数据的局部邻域结构，用于处理非线性特征关系。下图为聚类结果及与真实SSSA标签的对比图。

一張含有 文字, 螢幕擷取畫面, 圖表, 地圖 的圖片

自動產生的描述一張含有 文字, 地圖, 螢幕擷取畫面, 圖表 的圖片

自動產生的描述

图2 可视化t-SNE结果与K-means聚类标签以及可视化t-SNE结果与真实SSSA标签

# （2） 对比聚类结果和原本的SSSA标签。聚类结果与SSSA的分类结果是否相同？例如，同一簇内的运行状态是否都对应了稳定，是否都对应了不稳定？

从图1和图2可以看出PCA + KMeans 的调整兰德指数 (ARI) 为 0.0012，t-SNE + KMeans 的 ARI 为 0.0029，两者均接近 0，表明聚类结果与 SSSA 标签几乎无关。同时两者的轮廓系数分别为 0.43 和 0.38，说明簇内凝聚力有限，分界模糊。

虽然两种方法的聚类效果都一般，但对比之下，PCA + KMeans还是比t-SNE + KMeans更优胜一些。

无论是 PCA 还是 t-SNE 的降维，聚类结果（KMeans 标签）与原始 SSSA 标签的匹配程度都很低（ARI 接近 0），无法很好地对应 SSSA 的稳定和不稳定状态，原因可能是电力系统运行状态的数据可能存在高维复杂的非线性结构，简单的线性降维（如 PCA）或局部降维（如 t-SNE）未能有效捕捉这些关系，同时数据的分布过于复杂，导致 KMeans无法处理复杂的非球形簇。