k-means聚类实验

概述

- 利用*k*-means算法对LETTER数据集中的数据进行聚类。

数据说明

- LETTER数据集中的每一个类都对应一个英文字母。该数据集包含来自26个类，具有16个特征的20000个样例，每个类有大约769个样例。
- 数据集存储在data文件夹下。文件包括data和label字段，分别存储示例矩阵$X = \mathbb{R}^{N\times d}$和标记矩阵$Y = \mathbb{R}^{N}$。其中$N$是示例数量，$d$为特征维度，每个示例的标记$y \in \{1, 2, \ldots, 26\}$。

实验内容

- 利用欧式距离作为距离度量，在给定数据集上进行*k*-means聚类。
- 使用聚类性能度量的外部指标和内部指标对聚类结果进行分析。聚类度量的外部指标有Jaccard系数、FM指数、Rand指数等；聚类度量的内部指标有DB指数、Dunn指数等。
- （可选）实现其他聚类算法（如高斯混合聚类、层次聚类等），与*k*-means进行比较。
- 基于MindSpore平台提供的官方模型库，对相同的数据集进行训练，并与自己独立实现的算法对比结果（包括但不限于准确率、算法迭代收敛次数等指标），并分析结果中出现差异的可能原因。
- （加分项）使用MindSpore平台提供的相似任务数据集（例如，其他的分类任务数据集）测试自己独立实现的算法并与MindSpore平台上的官方实现算法进行对比，并进一步分析差异及其成因。

实验要求

- 推荐使用Python（在独立实现算法时，可采用Numpy，Pandas，Matplotlib等基础代码集成库；在使用MindSpore平台时，可使用平台提供的代码集成库）。
- 在独立实现算法时，不得使用集成度较高、函数调用式的代码库（如sklearn，PyTorch，Tensorflow等）。

作业提交格式

- 需要提供完整的可运行代码文件，聚类结果文件和实验报告，将以上内容打包压缩，压缩文件命名格式：学号-姓名-xxx实验。实验报告和代码注释应尽量详细。
- 实验报告内容参照报告模板，包括问题描述、实现步骤与流程、实验结果与分析、每个实验的心得体会（谈谈你自己的实现和MindSpore实现的差异、你在使用MindSpore平台过程中遇到的问题，以及想对平台改进提出的建议）、一个总的心得体会（谈一谈你对这门课程理论及实验的感悟与体会）。
- 代码和报告若有雷同，一律按0分处理。
- 若存在疑问，可以联系：seu_pr_2022@163.com