**深 圳 大 学 实 验 报 告**

|  |
| --- |
| **课程名称 机器学习**  **项目名称 实验五：K-means聚类算法**  **学 院 计算机与软件学院**  **专 业 软件工程（腾班）**  **指导教师 赖志辉**  **报 告 人 黄亮铭 学号 2022155028**  **实验时间 2024年4月22日至2024年5月12日**  **实验报告提交时间 2024年5月12日** |

**教务处制**

# 实验目的与要求

1. 熟悉聚类算法，要求理解聚类算法的基本原理。
2. 实现简单的聚类算法。
3. 使用实际数据集测试所写的聚类算法。
4. 分析聚类算法的效果。

# 二、实验内容与方法

1. 简述 K-means聚类算法的原理与算法过程.
2. 熟练掌握K-means聚类算法与结果的展示，并代码实现，做一个2维或三维空间中的2~3类点（每个类有10个点）聚类实验，把聚类结果用不同的颜色与符号表示。
3. 实现人脸图像（取前2~3个人的人脸图像）聚类实验与旋转物体（在COIL20数据集中取前2~3个类的图像）， 把聚类结果用不同的颜色与符号表示，并把对应的图像放在相应点的旁边，让人一眼看出结果对不对；同时列表给出其在不同数据库在不同K时的聚类精度。
4. 把课堂上我们讲课时说的模型及其优化过程写在实验最后的思考题部分。

# 三、实验步骤与过程

1. 简述 K-means聚类算法的原理与算法过程.

**聚类的定义** 假定样本集包含m个无标记样本，每个样本是一个n维特征向量，则聚类算法将样本划分为k个不相交的簇，其中相应地，我们可以使用来表示样本簇标记。由此聚类的结果可以表示为。

**K-means聚类算法简介** k-均值聚类是一种常用的聚类算法，用于将数据集中的样本分成 k 个类别。它的原理是将样本分配给 k 个类别，使得每个样本都属于与其最近的类别中心点所代表的类别。然后，计算每个类别的新的中心点，重复这个过程直到类别的中心点不再发生变化或达到预定的迭代次数。

**K-means聚类算法原理** 给定样本集，k均值算法针对聚类所得簇划分最小化平方误差

其中簇均值向量。

**K-means聚类算法过程** Step1：读入数据集，并进行相应的预处理。Step2：设置簇类数目k，设置迭代允许误差，设置迭代次数；随机选取k个样本作为初始均值向量。Step3：根据公式计算样本与均值向量之间的欧氏距离，根据距离最近的均值向量确定簇标记，将当前样本划归到相应的簇中.Step4：判断是否满足迭代停止条件，如不满足，返回Step3继续迭代；如满足，退出迭代，并输出相应的簇划分。

1. 熟练掌握K-means聚类算法与结果的展示，并代码实现，做一个2维或三维空间中的2~3类点（每个类有10个点）聚类实验，把聚类结果用不同的颜色与符号表示。

**K-means聚类算法代码实现**

* + 1. 首先读入数据集。

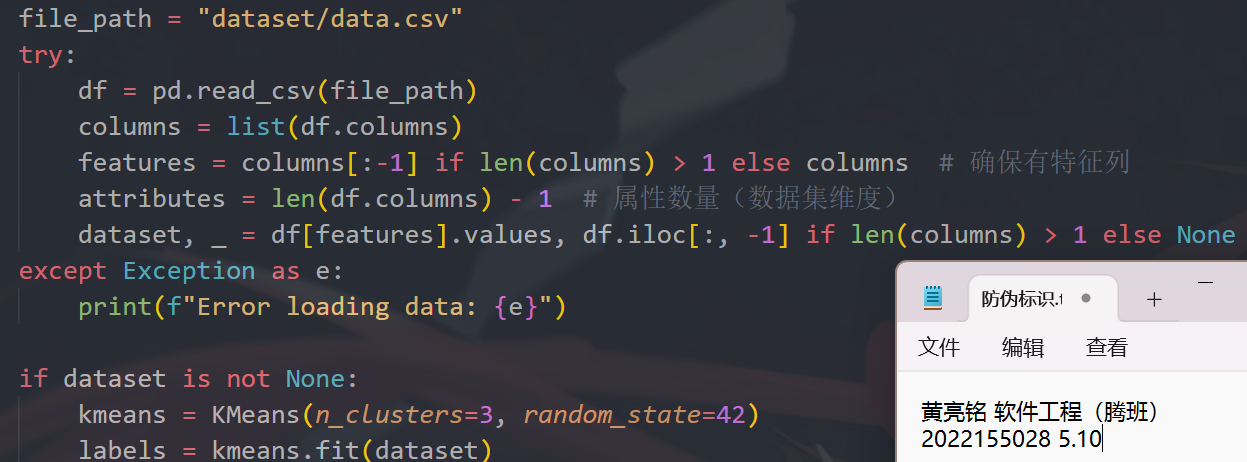


图1：读入数据集

* + 1. 然后初始化聚类。



图2：初始化聚类

* + 1. 再执行K-means聚类算法的迭代过程。

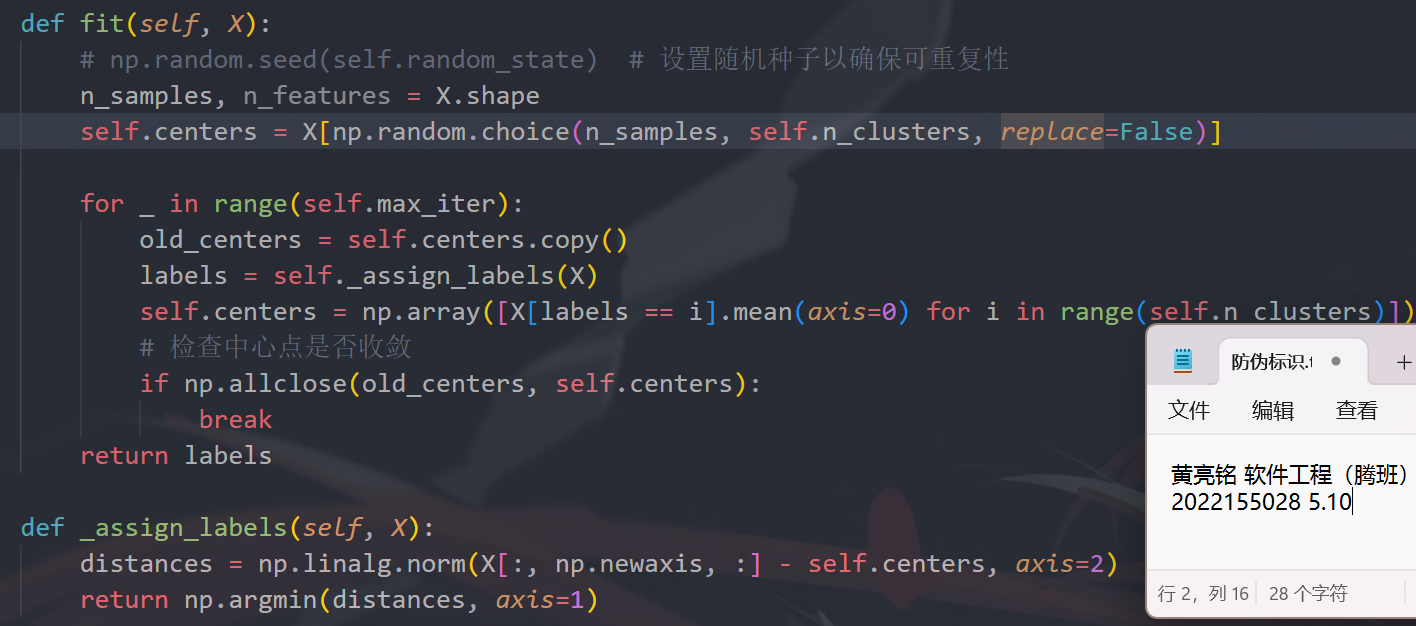


图3：迭代过程

* + 1. 输出可视化结果。



图4：输出可视化结果

**K-means聚类算法可视化结果展示。**

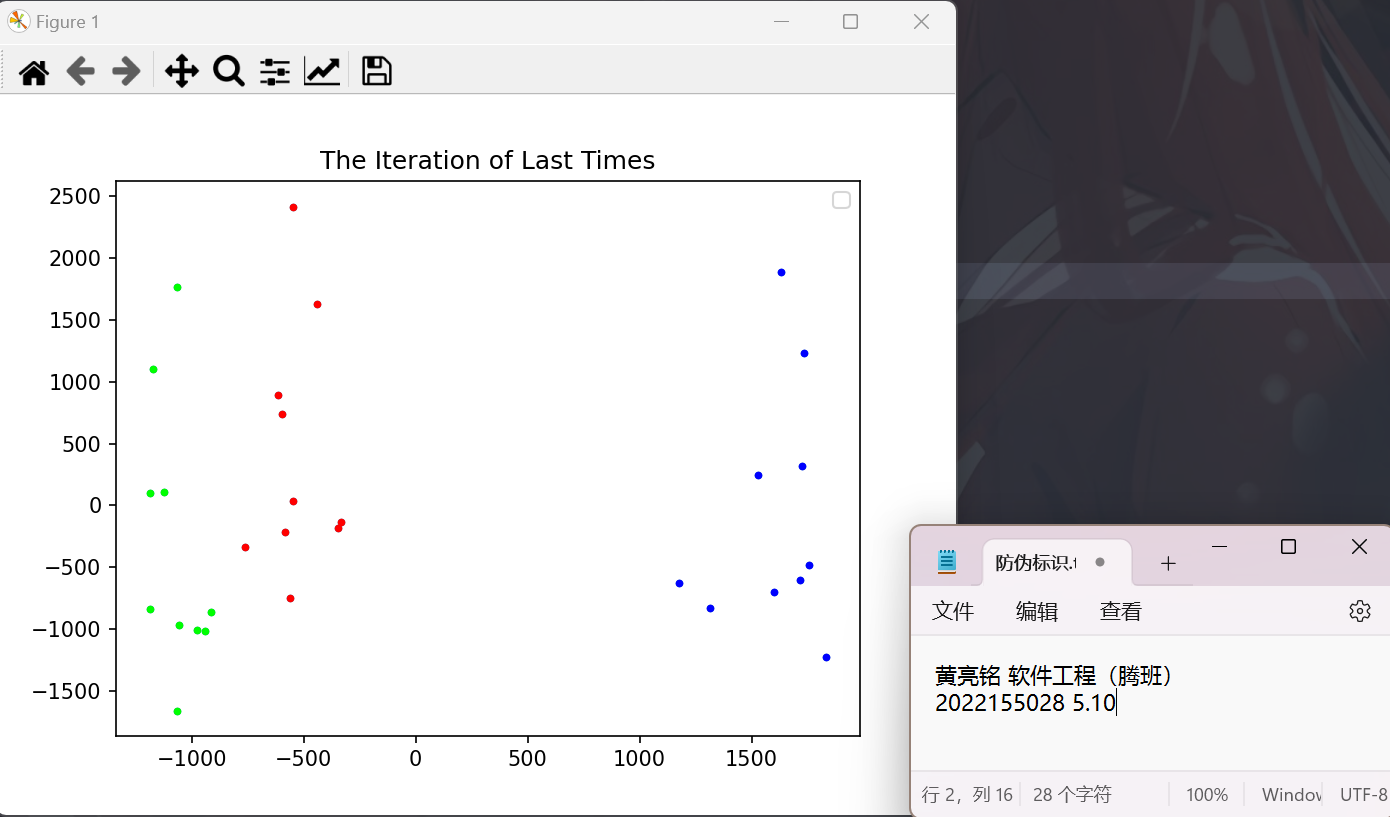


图5：可视化结果

1. 实现人脸图像（取前2~3个人的人脸图像）聚类实验与旋转物体（在COIL20数据集中取前2~3个类的图像）， 把聚类结果用不同的颜色与符号表示，并把对应的图像放在相应点的旁边，让人一眼看出结果对不对；同时列表给出其在不同数据库在不同K时的聚类精度。

**人脸图像的聚类结果（K=3）。**

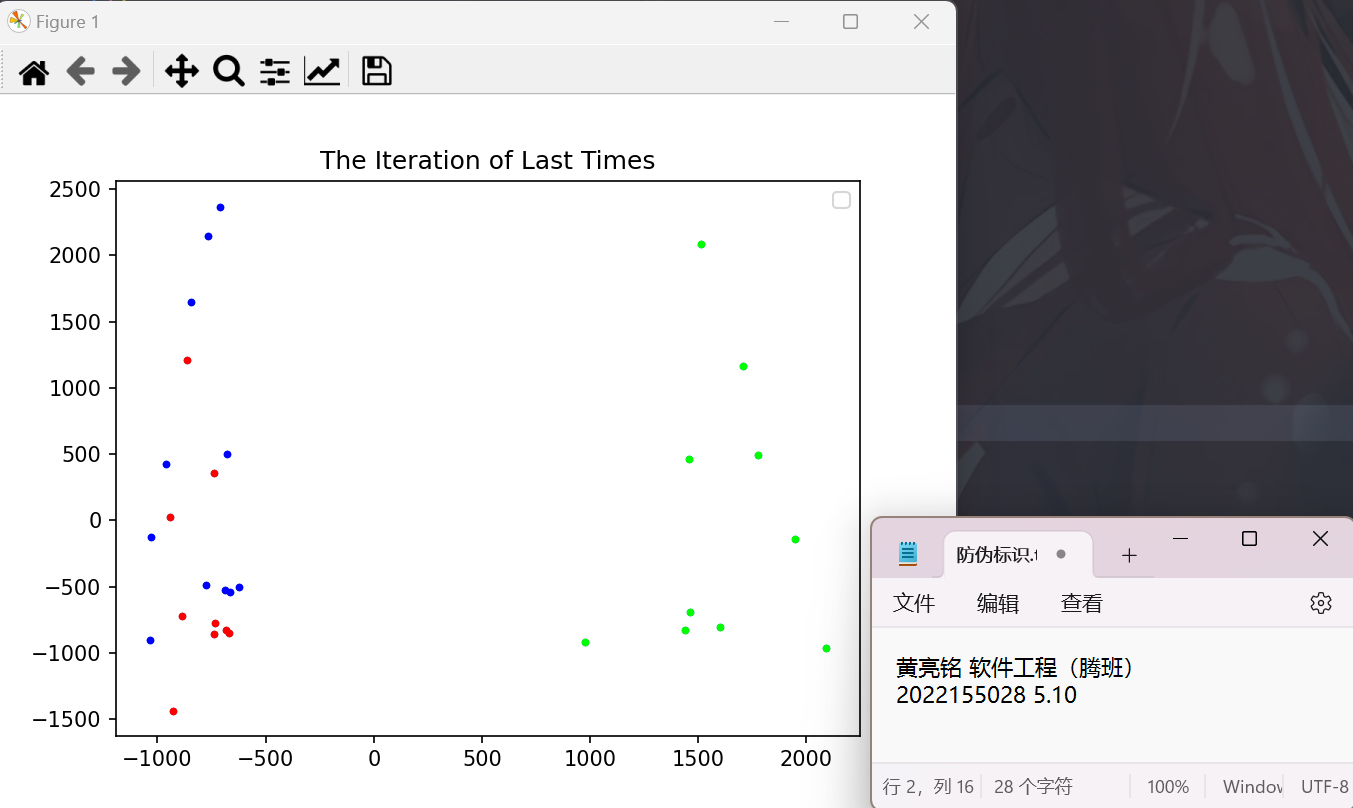


图6：二维空间的人脸图像的聚类结果

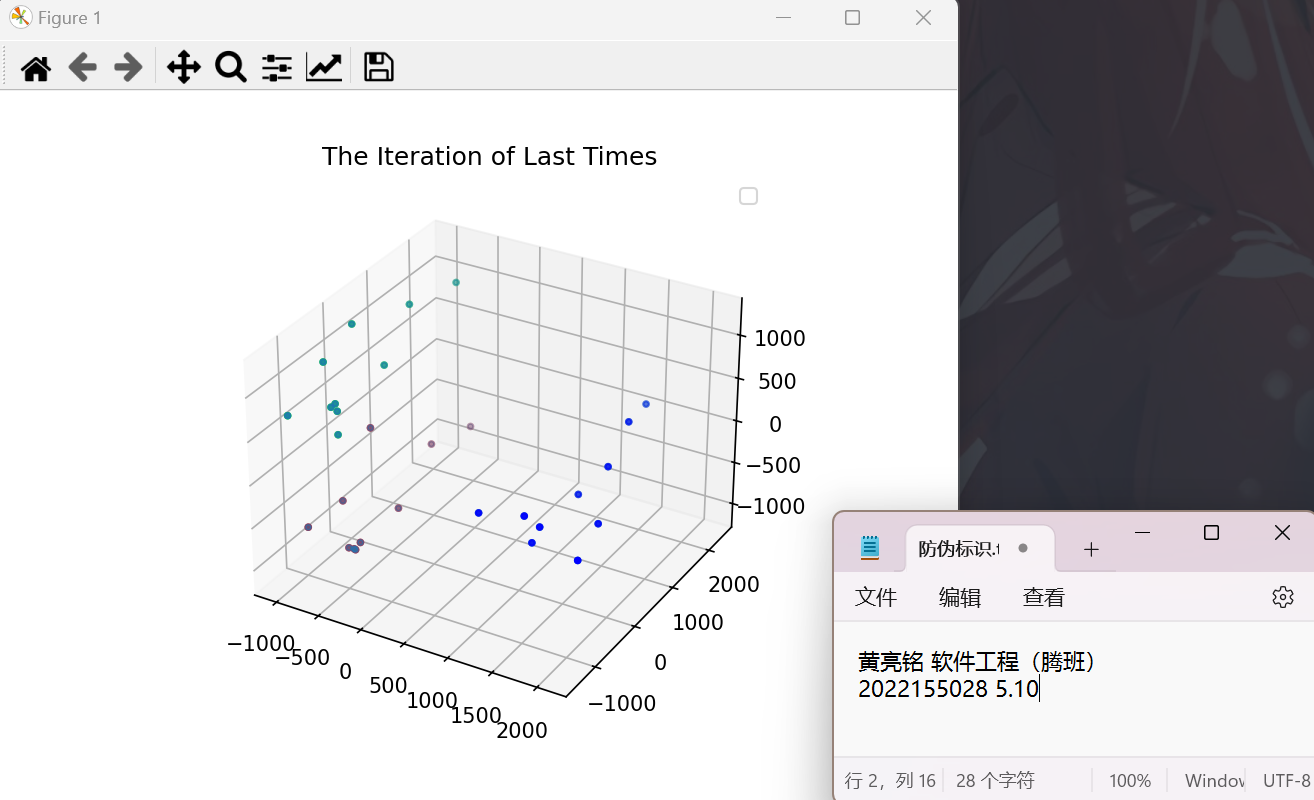


图7：三维空间的人脸图像的聚类结果

**人脸图像的聚类结果分析。**

观察二维和三维的人脸图像聚类结果，发现在二维空间中，样本点的簇划分并不合理。但是在三维空间中，发现样本点的簇划分比较合理。出现这种情况的原因可能是即使特征值最大的两个特征向量代表的特征在部分人脸图像中也高度相似，因此出现了不同簇的样本点交叉重合的现象。而在三维空间中增加了一个特征向量，使样本点可视化的簇划分更加合理。

**旋转物体的聚类结果（K=3）。**

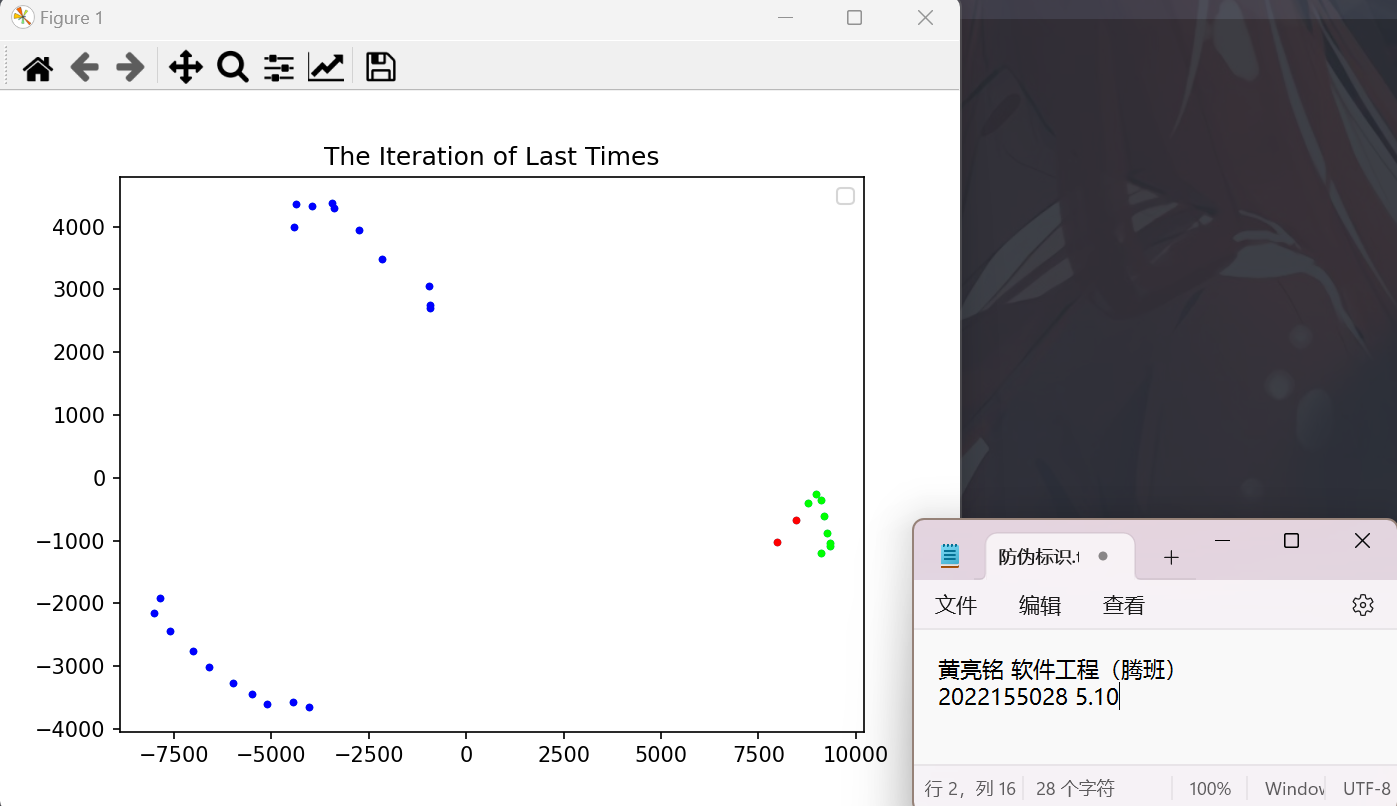
****

图8：二维空间的旋转物体的聚类结果

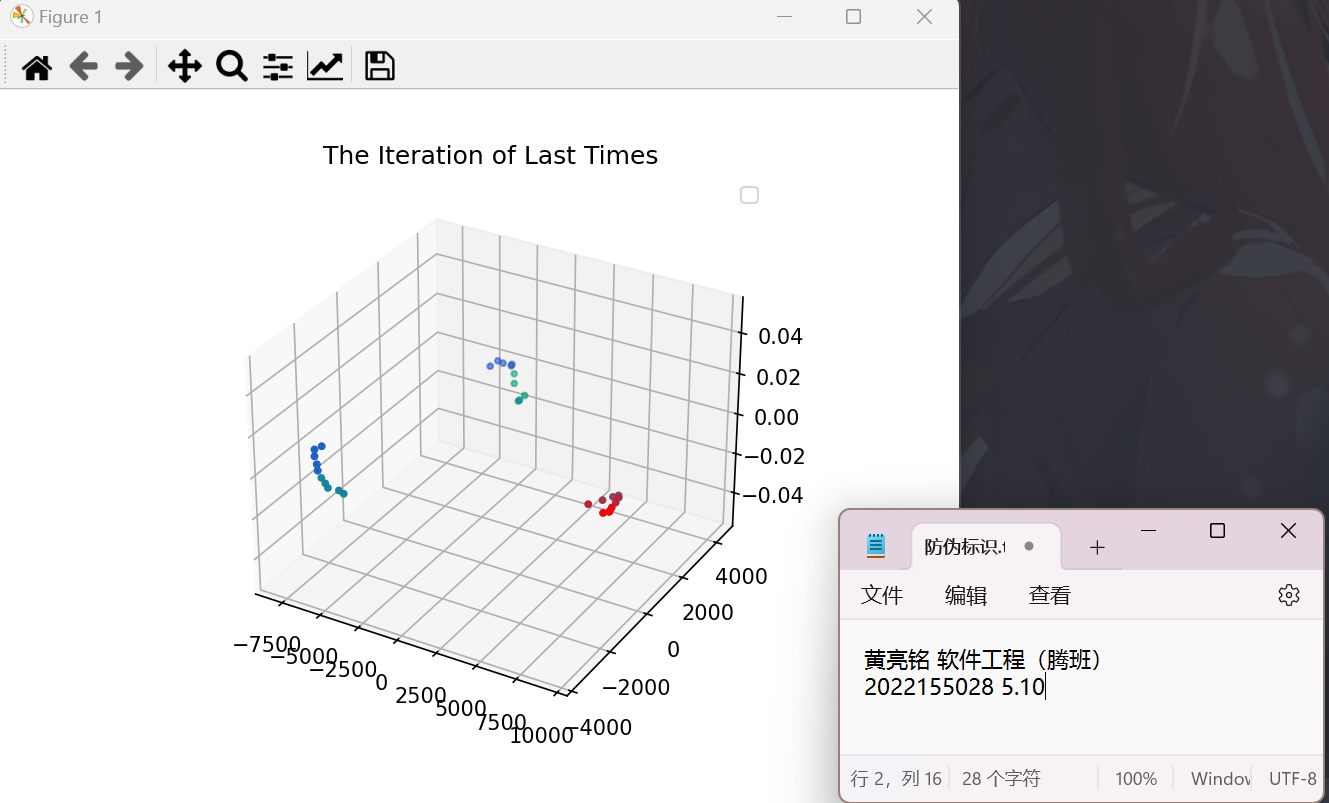
****

图9：三维空间的旋转物体的聚类结果

**旋转物体的聚类结果分析。**

观察二维和三维的旋转图像聚类结果，发现无论是在二维空间还是在三维空间中，K-means聚类算法对旋转物体的划分都不是很理想，预测标签和真实标签的相似率即预测正确率仅有70%。这种情况的出现可能是由于K-means聚类算法难以处理旋转物体。

**不同数据库在不同K值时的聚类精度。**

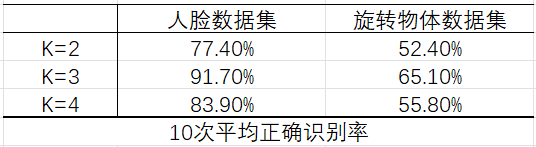
****

图10：不同数据集在不同K值的聚类精度

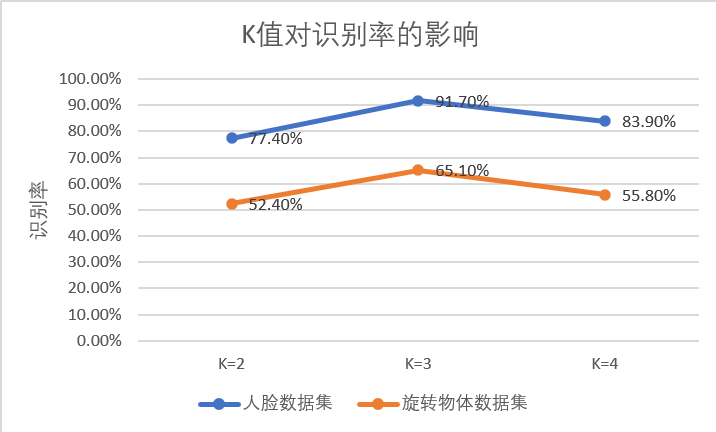
****

图11：不同数据在不同K值的聚类精度可视化

1. 使用其他聚类方法，并且和K-means聚类算法比较。

**模糊C均值聚类算法（下称FCM聚类算法）**简介 FCM聚类是一种基于模糊逻辑的聚类算法，它允许每个样本点以一定的隶属度属于每个聚类中心。在FCM中，每个样本点与每个聚类中心之间的关系不是二元的（属于或不属于），而是以一个在0到1之间的隶属度值来表示。这个值表示了样本点属于该聚类的程度，隶属度值越高表示样本点与该聚类的关联性越强。FCM的目标是通过最小化一个目标函数来优化聚类中心和样本点的隶属度，这个目标函数定义了样本点与聚类中心之间的加权欧氏距离的平方和，其中权重是样本点与聚类中心之间的隶属度。通过迭代更新聚类中心和样本点的隶属度，FCM算法寻找到一个局部最优的聚类结果。

**FCM聚类算法定义** 给定一个数据集,其中每个样本点都属于聚类中心（）之一。对于每个样本点和每个聚类中心定义隶属度，表示样本点属于聚类中心的程度。

隶属度满足的条件：1）为非负实数；2）对于每个样本点，隶属度之和等于1，即；3）隶属度表示样本点之间的关联程度，值越大表示关联程度越高。

**FCM聚类算法的目标**

**FCM聚类算法的目标函数求解过程**

, ,

使用拉格朗日乘子法：

分别对和求偏导，得:

其中，为欧几里得距离。

令， ，得： ，由此得到

**FCM聚类算法的目标函数求解结果。**

**FCM聚类算法的流程。**

算法输入：簇类数c，数据集，模糊指标，终止误差。

算法输出：簇类中心向量矩阵和隶属度矩阵

具体操作步骤：

Step1：读入数据集，并进行相应的预处理。

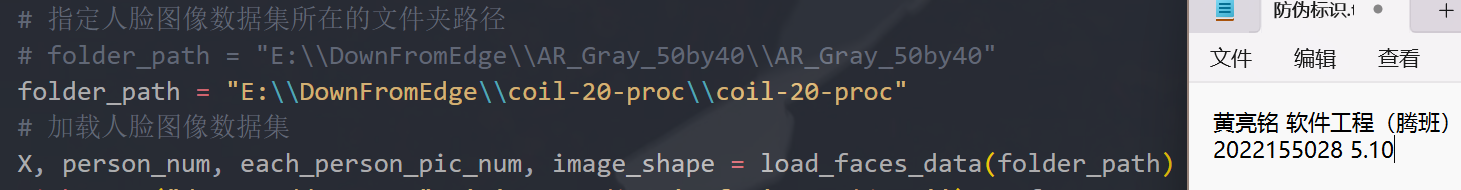


图12：读入数据集

Step2：设置簇类数目c；设置模糊指标，设置迭代允许误差，设置迭代次数；初始化隶属度矩阵。



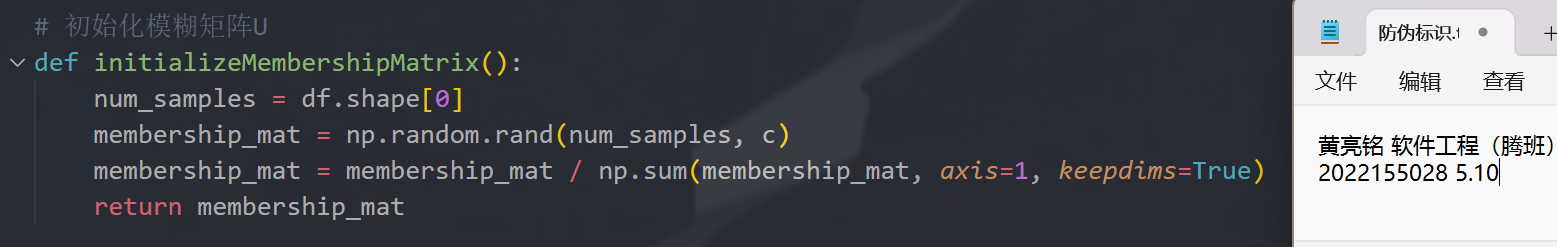


图13：初始化

Step3：从样本中随机抽取个样本作为初始簇类中心。

Step4：根据公式更新隶属度矩阵和聚类中心。





图14：更新隶属度矩阵和聚类中心

Step5：判断是否满足迭代停止条件，如不满足，返回Step3继续迭代；如满足，退出迭代，并输出簇类中心向量矩阵和隶属度矩阵。

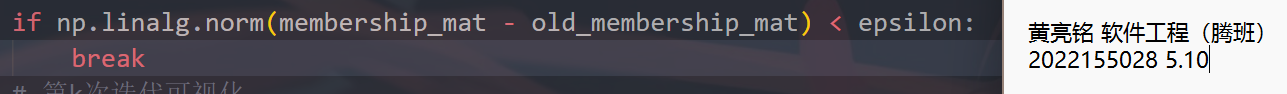


图15：判断迭代是否终止

在人脸数据集中的可视化结果。

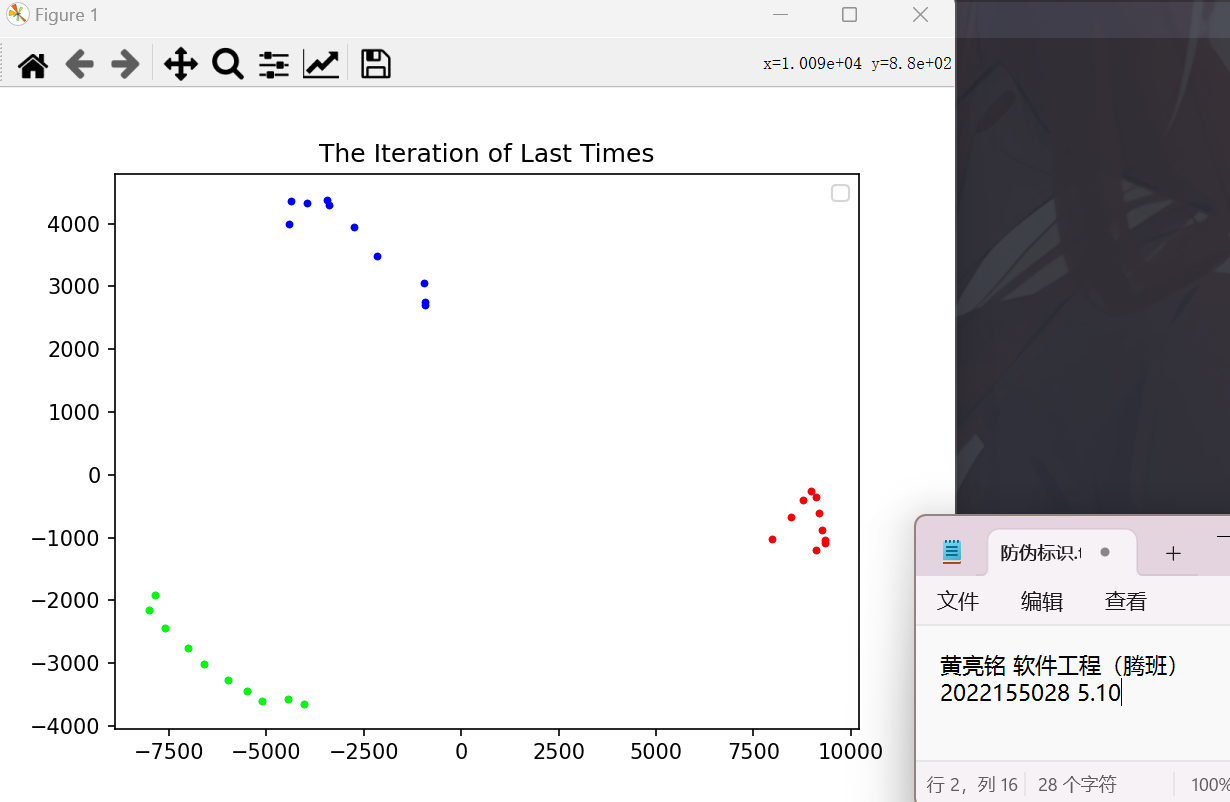


图16：二维可视化结果

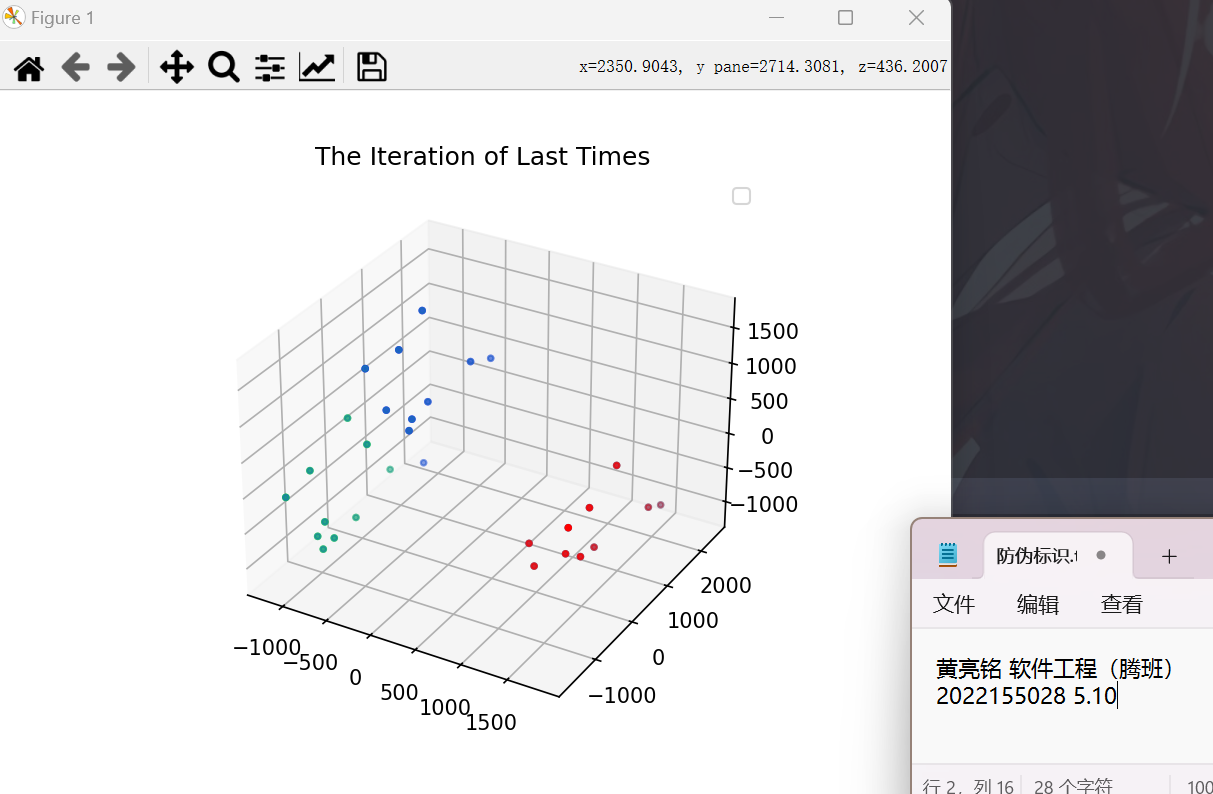


图17：三维可视化结果

在旋转数据集中的可视化结果。

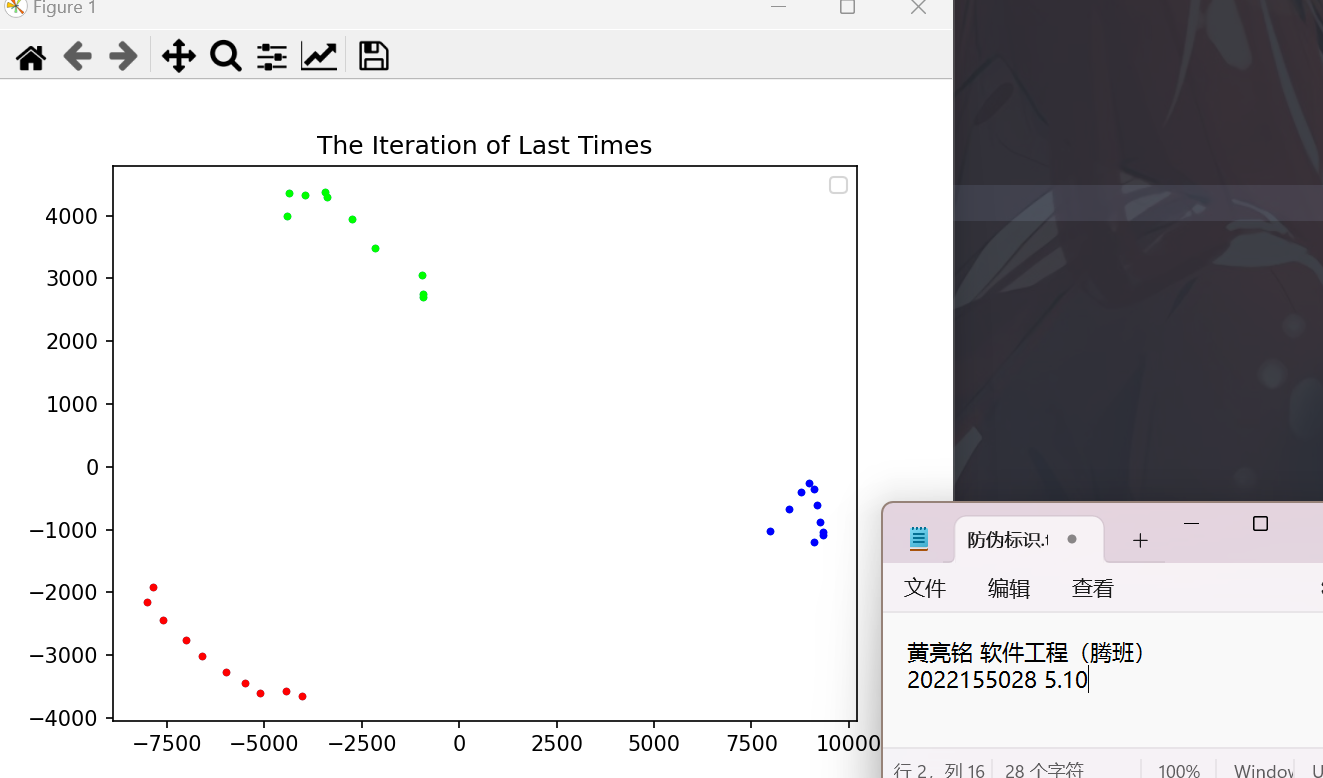


图18：二维可视化结果

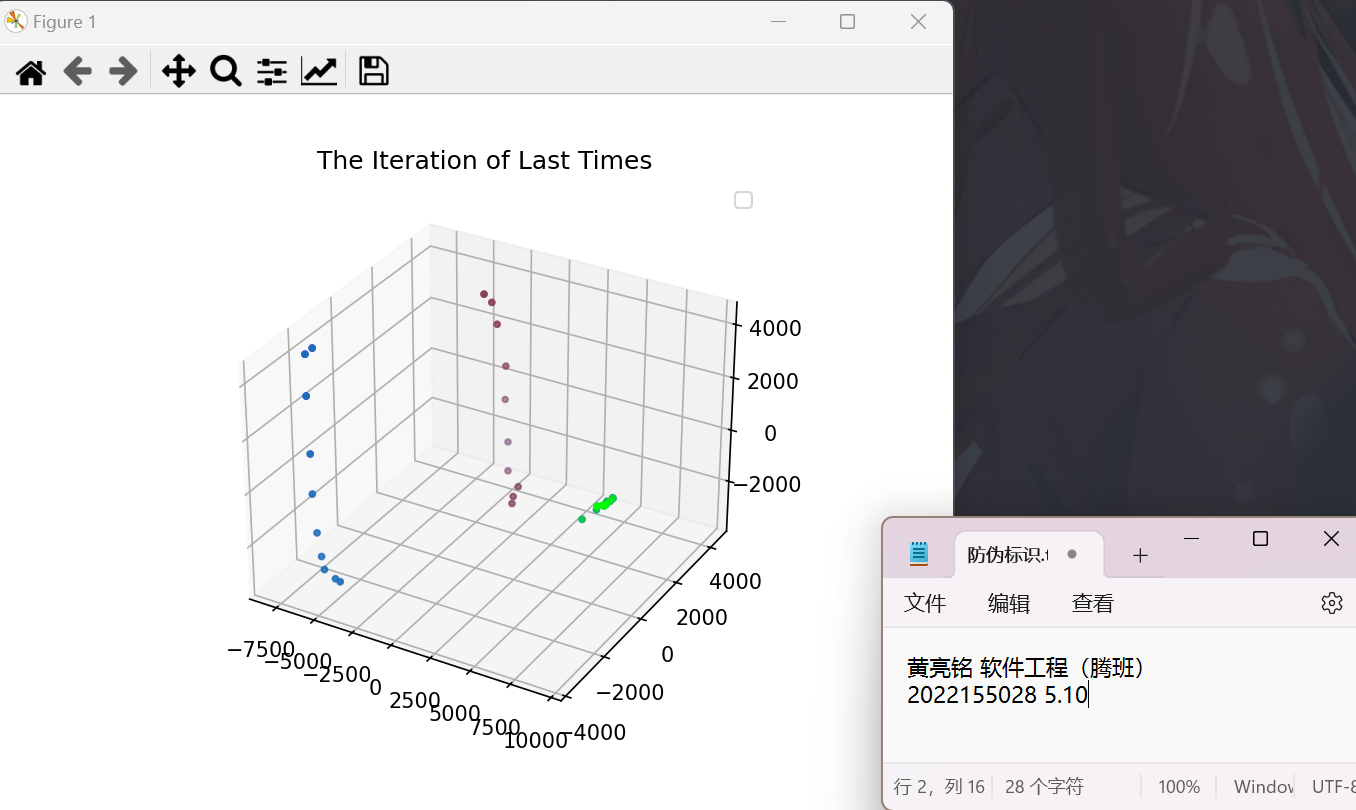


图19：三维可视化结果

1. 不同聚类算法在K=3时在不同数据集的聚类精度



图20：不同聚类算法在K=3时在不同数据集的聚类精度

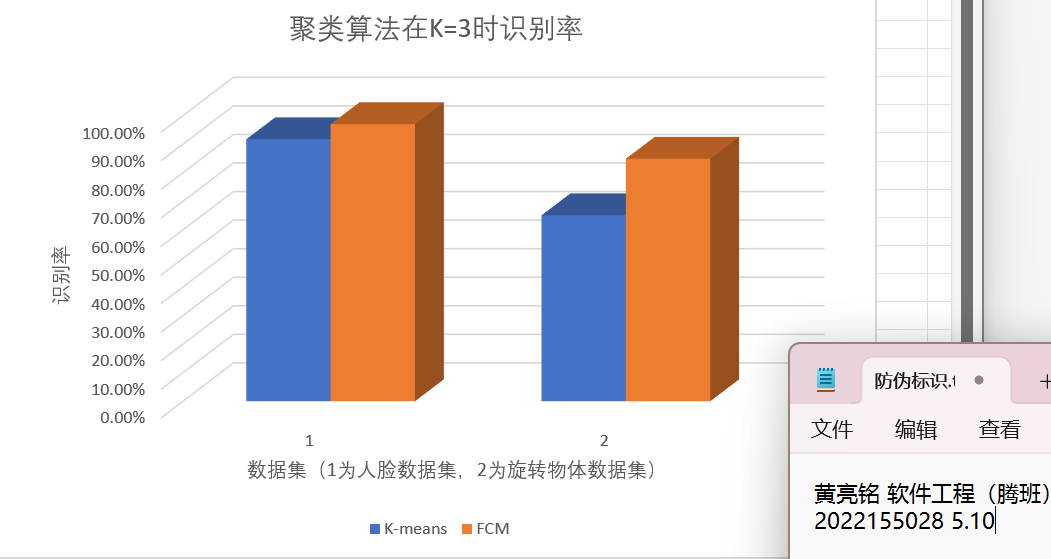
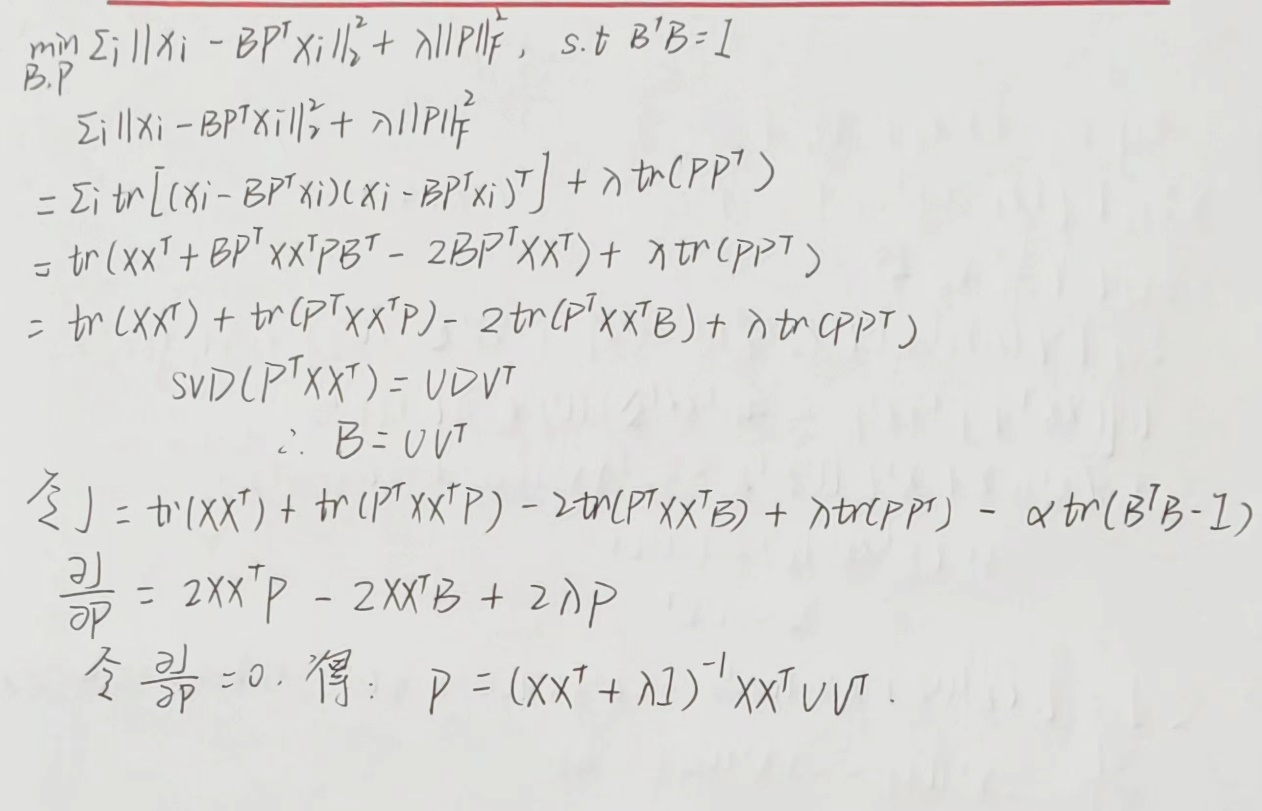


图21：不同聚类算法在K=3时在不同数据集的聚类精度可视化

# 四、实验结论或体会

1. K-means聚类算法在处理人脸图像数据集时表现良好，但在处理旋转物体数据集时表现较差。
2. FCM聚类算法在人脸数据集和旋转物体数据集中的聚类精度对比K-means聚类算法均有提升，特别是对于旋转物体数据集的聚类精度，有较大的提升，但是识别率仍然相对较低。
3. 通过本次实验，我熟悉了聚类算法，理解了聚类算法的基本原理。
4. 通过本次实验，我实现了聚类算法的代码，更加深刻地了解了聚类算法。

# 五、思考题



思考题1



思考题2

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。