上海理工大学光电信息与计算机工程学院

**《数据挖掘》实验报告**

****

**专　　业 智能科学与技术**

**姓 名　　高浩琦**

**学　 号 2035060413**

**年　　级 2020级**

**指导教师**

**成 绩：**

**教师签字：**

# 实验四 K-means聚类

一、实验目的和要求

1，掌握基于距离的K-means聚类的计算过程

2，通过Python编程还原计算过程，对数据进行聚类

3，通过运行程序更改不同K值对同一数据进行聚类，理解K-means聚类中K均值初始值选择对聚类造成的差异

二、实验算法

import numpy as np

import PIL.Image as image

from sklearn.cluster import KMeans

def loadData(filePath):

    data = []

    img = image.open(filePath)

    m,n = img.size

    for i in range(m):

        for j in range(n):

            x,y,z = img.getpixel((i,j))

            data.append([x/256.0,y/256.0,z/256.0])

    return  np.mat(data),m,n

imgData,row,col = loadData('block.jpg')

label = KMeans(n\_clusters=4).fit\_predict(imgData)

label = label.reshape([row,col])

pic\_new = image.new("L",(row,col))

for i in range(row):

    for j in range(col):

        pic\_new.putpixel((i,j),int(256/(label[i][j]+1)))

pic\_new.save("result-bull-4.jpg","JPEG")

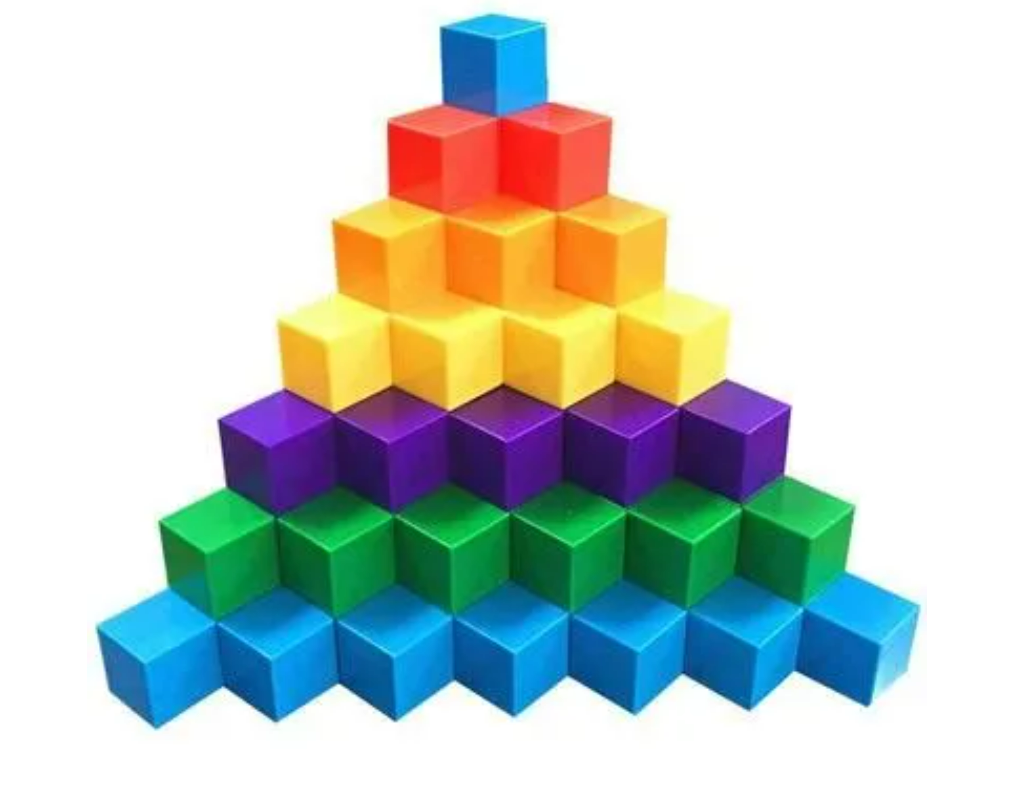
三、实验过程记录

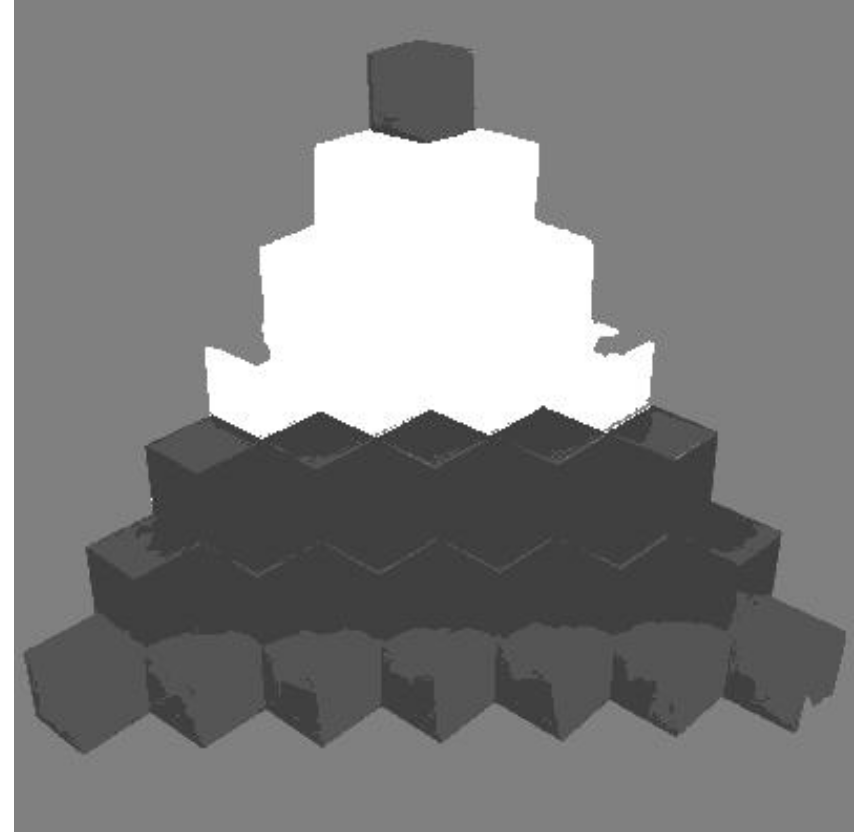
1．读取图片数据

2．利用kmeans算法计算图片

3. 输出图片

四、结果与分析





五、自评

Kmeans算法是最常用的聚类算法，主要思想是:在给定K值和K个初始类簇中心点的情况下，把每个点(亦即数据记录)分到离其最近的类簇中心点所代表的类簇中，所有点分配完毕之后，根据一个类簇内的所有点重新计算该类簇的中心点(取平均值)，然后再迭代的进行分配点和更新类簇中心点的步骤，直至类簇中心点的变化很小，或者达到指定的迭代次数。

通过本次实验 ，对kmeans算法的理解大大加深了。