数 据 挖 掘

课 程 设 计 报 告

姓名： 班级： 学号：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 实验名称 | 规范程度 | 原理叙述 | 实验过程 | 实验结果 | 实验成绩 |
| 必选 | 主成分分析 |  |  |  |  |  |
| 支持向量机 |  |  |  |  |  |
| K均值聚类 |  |  |  |  |  |
| 可选  (七选三) | 决策树 |  |  |  |  |  |
| BP神经网络 |  |  |  |  |  |
| 朴素贝叶斯分类 |  |  |  |  |  |
| 奇异值分解 |  |  |  |  |  |
| PageRank |  |  |  |  |  |
| K近邻分类 |  |  |  |  |  |
| EM算法 |  |  |  |  |  |
| 成绩 | 平均成绩 |  | | | | |
| 折合成绩 |  | | | | |

注：1、每个实验中各项成绩按照5分制评定，实验成绩为各项总和。

2、平均成绩取各项实验平均成绩。

3、折合成绩按照教学大纲要求的百分比进行折合。

2021年 7月

**课程设计报告要求：**

1. **书写格式**：标题采用“宋体小四号”，正文使用“宋体五号”，打印页码，页眉页脚使用第3页参考格式。

2. **内容**: 每个实验不要超过5页，最后一页总结，正反面打印。一份实验报告不要超过15张纸（即30页）。

3. **装订和打印**：封面为第一页，无需打印封底。按照上面的格式采用A4纸打印。装订时：订书钉在报告左边沿装订三个，上中下各一个（对齐）。电子版以学号姓名命名。

4．**交课程设计报告：一人一份**。7月10日前先交至班长或学习委员处，纸质版收齐交至东8206办公室，电子版以班级为单位拷贝给曾山老师。

### 示例：K均值聚类算法

**一、实验目的**

将模式识别方法与图像处理技术相结合，掌握利用K均值聚类算法进行图像分类的基本方法，通过实验加深对基本概念的理解。

**二、实验仪器设备及软件**

MATLAB、WIT、Python、Java

**三、实验原理**

K均值聚类法分为三个步骤：

初始化聚类中心：

1、根据具体问题，凭经验从样本集中选出C个比较合适的样本作为初始聚类中心。

2、用前C个样本作为初始聚类中心。

3、将全部样本随机地分成C类，计算每类的样本均值，将样本均值作为初始聚类中心。

初始聚类：

1、按就近原则将样本归入各聚类中心所代表的类中。

2、取一样本，将其归入与其最近的聚类中心的那一类中，重新计算样本均值，更新聚类中心。然后取下一样本，重复操作，直至所有样本归入相应类中。

判断聚类是否合理：

1. 采用误差平方和准则函数判断聚类是否合理，不合理则修改分类。循环进行判断、修改直至达到算法终止条件。
2. 聚类准则函数

误差平方和准则函数（最小平方差划分）

 (1)

 (2)

1. 单样本改进：每调整一个样本的类别就重新计算一次聚类的中心

 (3)

只调整一个样本。

**四、实验步骤及程序**

**1、实验步骤**

理解K均值算法基本原理，编写程序实现对自选图像的分类，并将所得结果与WIT处理结果进行对比。

K均值算法步骤：

1. 给定类别数C和允许误差，；
2. 初始化聚类中心；
3. 修正，0其他，；
4. 修正聚类中心；
5. 计算误差如果则结束，否则转（3）。

**2、实验源程序**

clc

clear

tic

RGB= imread ('Water lilies.jpg');

img=rgb2gray(RGB);

[m,n]=size(img);

subplot(2,2,1),imshow(img);title('原图像')

subplot(2,2,2),imhist(img);title('原图像的灰度直方图')

hold off;

img=double(img);

for i=1:200

c1(1)=25;

c2(1)=125;

c3(1)=200;

r=abs(img-c1(i));

g=abs(img-c2(i));

b=abs(img-c3(i));

r\_g=r-g;

g\_b=g-b;

r\_b=r-b;

n\_r=find(r\_g<=0&r\_b<=0);

n\_g=find(r\_g>0&g\_b<=0);

n\_b=find(g\_b>0&r\_b>0);

i=i+1;

c1(i)=sum(img(n\_r))/length(n\_r);

c2(i)=sum(img(n\_g))/length(n\_g);

c3(i)=sum(img(n\_b))/length(n\_b);

d1(i)=abs(c1(i)-c1(i-1));

d2(i)=abs(c2(i)-c2(i-1));

d3(i)=abs(c3(i)-c3(i-1));

if d1(i)<=0.001&&d2(i)<=0.001&&d3(i)<=0.001

R=c1(i);

G=c2(i);

B=c3(i);

k=i;

break;

end

end

img=uint8(img);

img(find(img<R))=0;

img(find(img>R&img<G))=128;

img(find(img>G))=255;

toc

subplot(2,2,3),imshow(img);title('聚类后的图像')

subplot(2,2,4),imhist(img);title('聚类后的图像直方图')

**五、实验结果与分析**

**1.WIT结果**

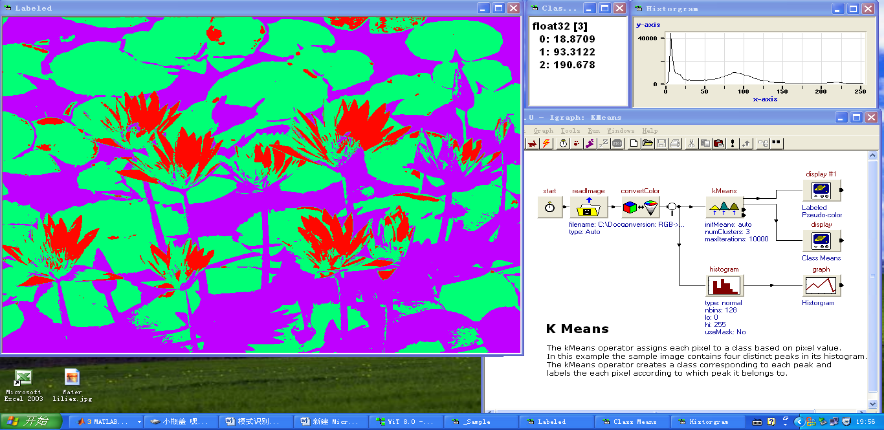
****

图1 WIT聚类分析系统分析界面

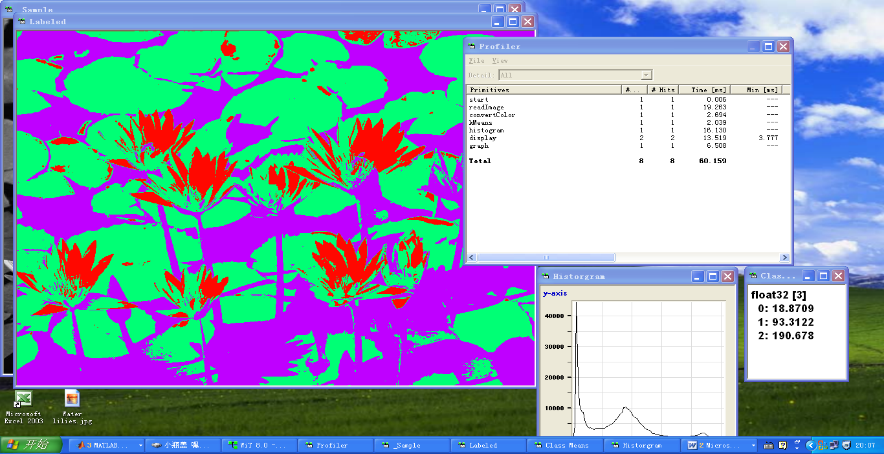
****

图2 WIT聚类分析系统分析结果

聚类类别数 3

聚类中心 R=18.8709 G=93.3122 B=190.678

迭代次数 256

运行时间 60.159ms

**2、K均值聚类结果**

****

图5 K均值聚类分析结果

聚类类别数 3

聚类中心 R =19.9483 G =94.4184 B =192.3320

迭代次数 8

运行时间 2.278493 seconds

K均值聚类方法和WIT系统操作后对应的聚类中心误差较小，分别是19.9483 94.4184 192.3320和 18.8709 93.3122 190.678。说明K均值聚类分析方法误差较小，但是相较于WIT系统分析的时间，K均值聚类时间较长，迭代次数较多。