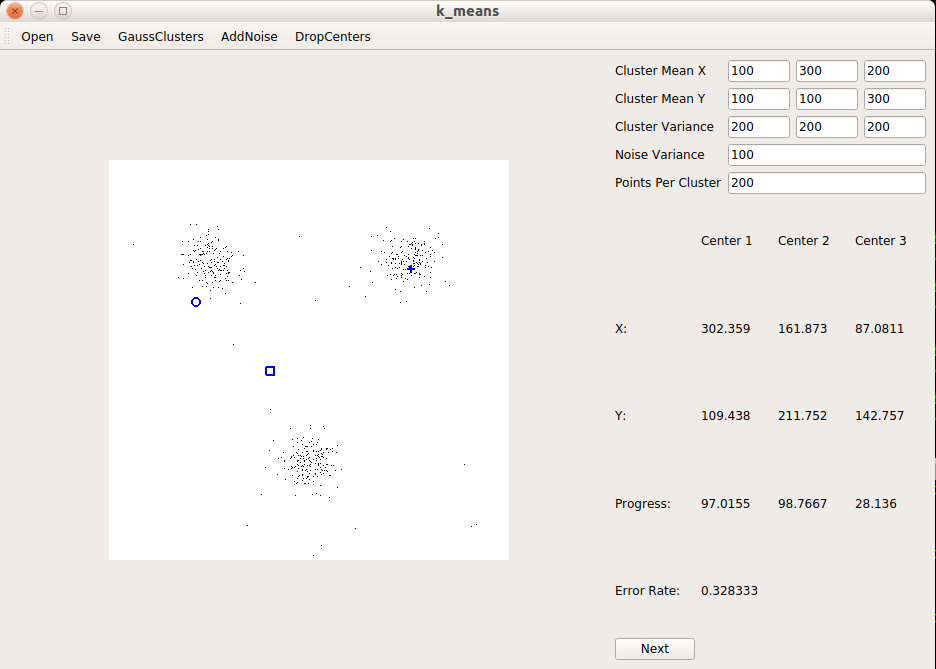
# K-means实验报告

## 仿真程序简介

本仿真程序算法部分用C++完成，使用了一些C++14新增的特性。界面部分调用Qt库提供的GUI控件完成。



工具栏中Open，Save能浏览或保存RAW格式的文件。与本实验无关。GaussClusters用来生成服从高斯分布的点集合。AddNoise用来为已经生成的高斯点集添加噪声。DropCenters用来随机生成中心点的初始位置。界面主体中，左半部分是观察窗，能够观察到所生成点集的形态以及中心点的位置。右上角是调整参数的交互控件，可以调整服从高斯分布点集的期望，方差，所加噪声的方差以及每个点集内部含有的点数。从右下角可以持续观察中心收敛的过程。可以观察到中心的位置，与上一位置之间的偏移距离，以及当前的分类误差率。

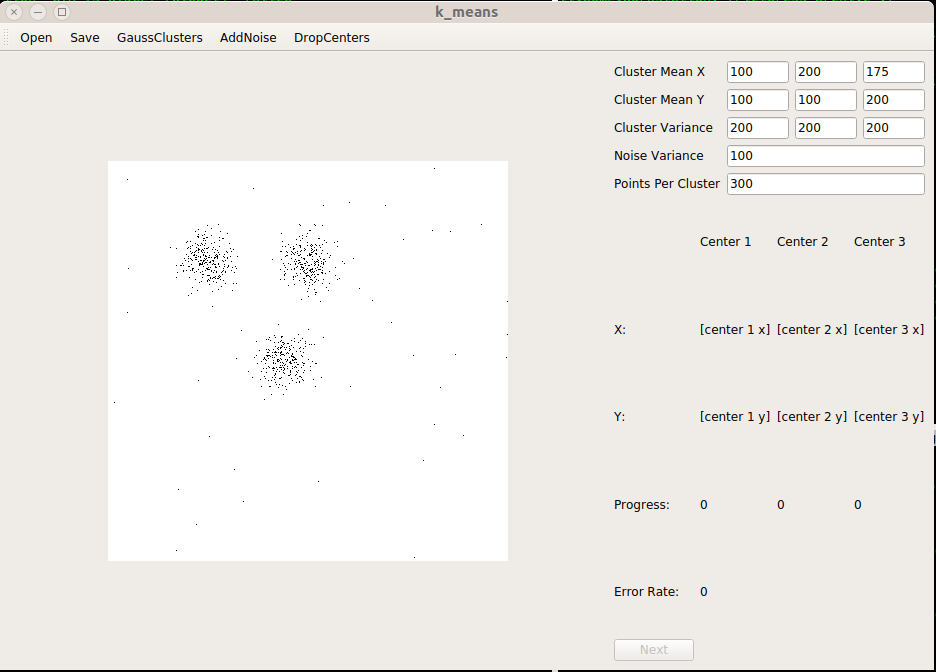
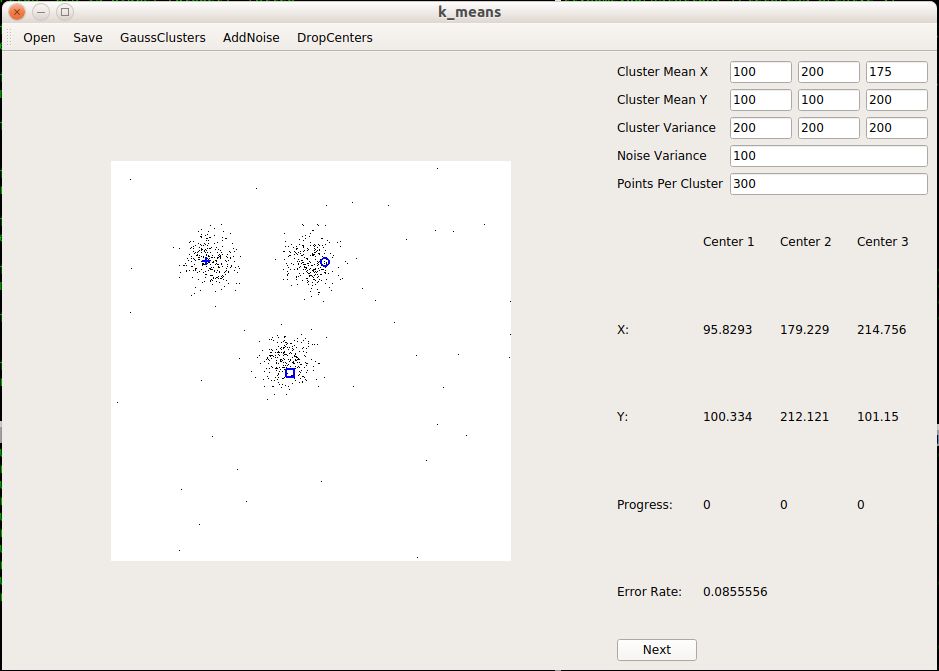
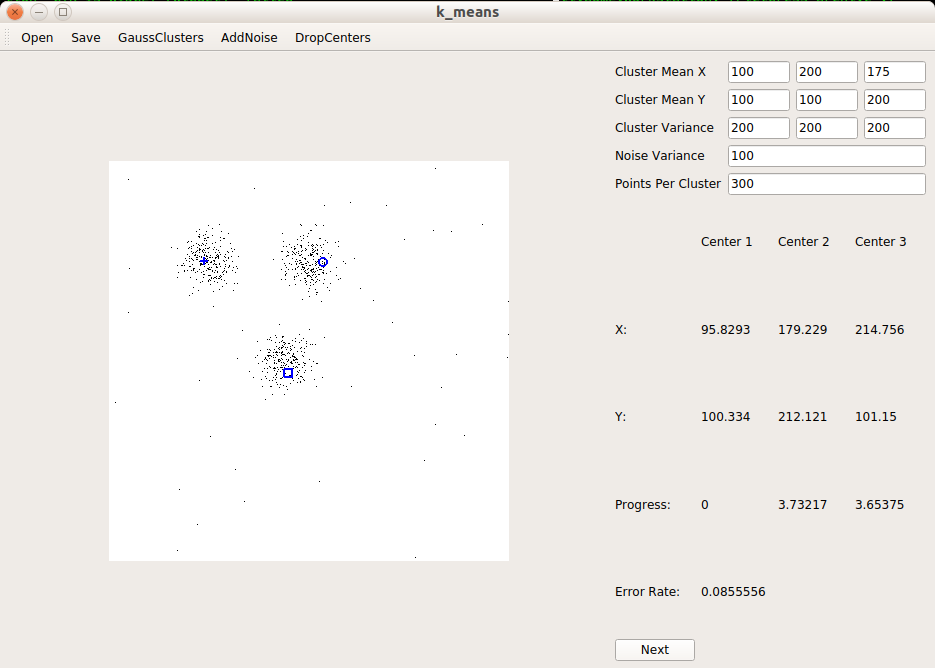
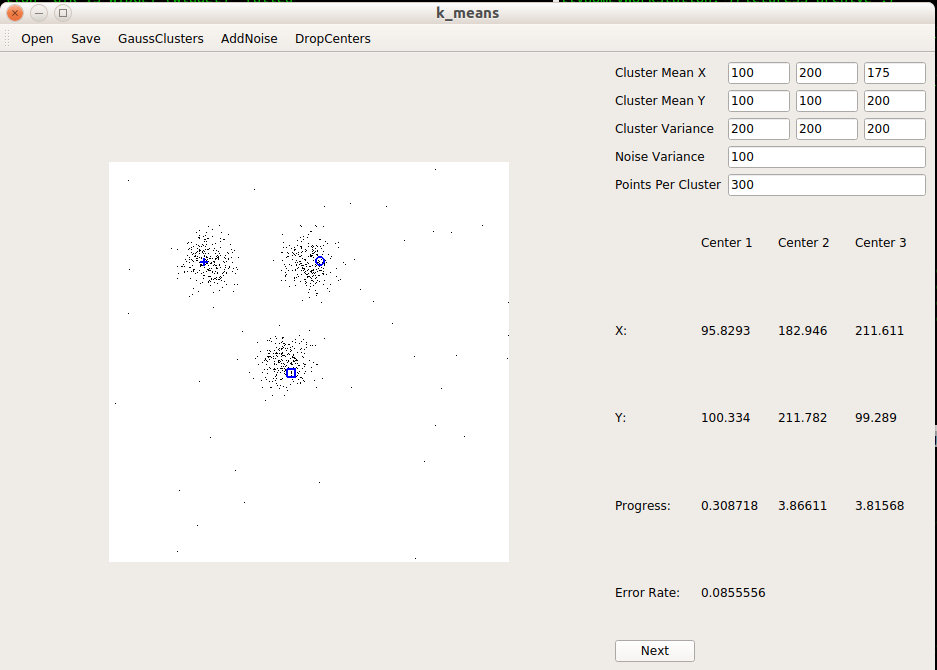
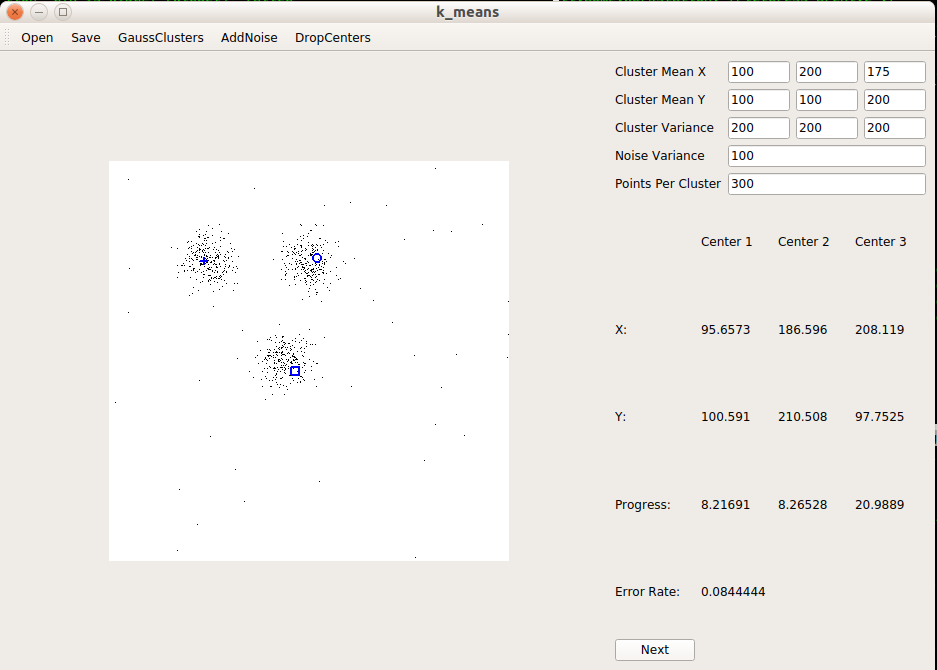
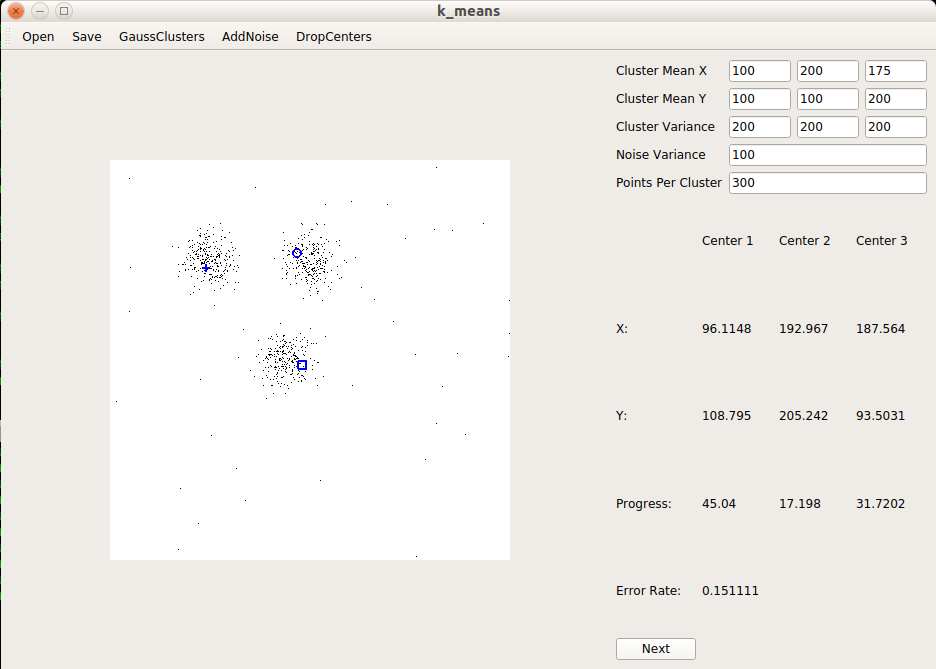
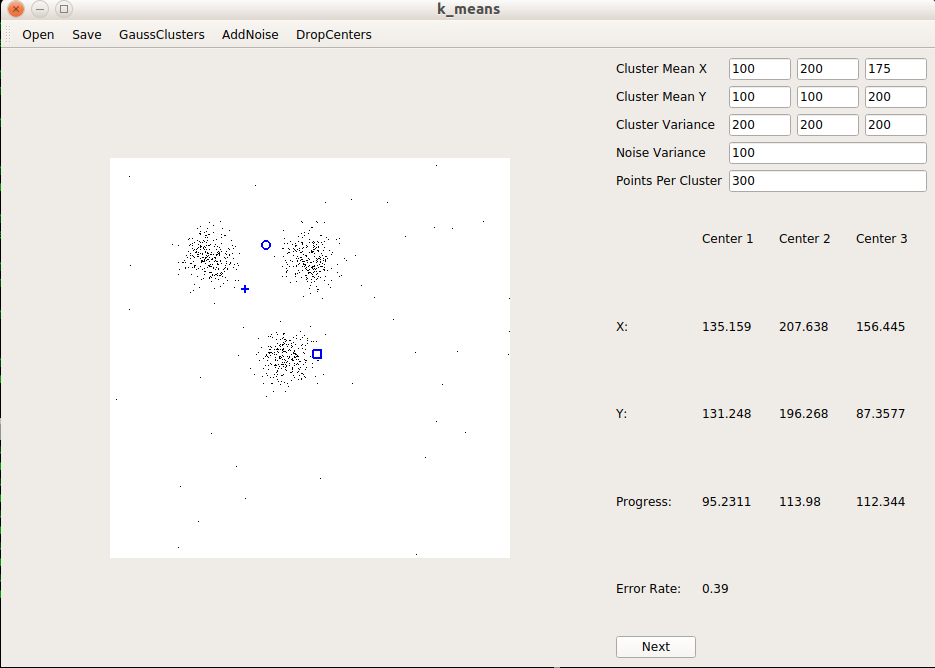
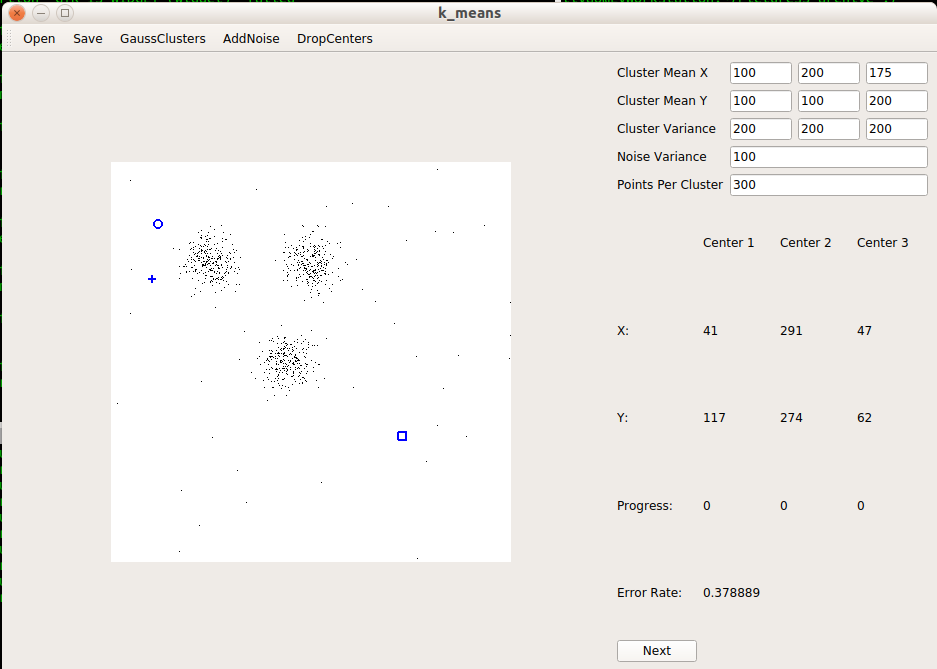
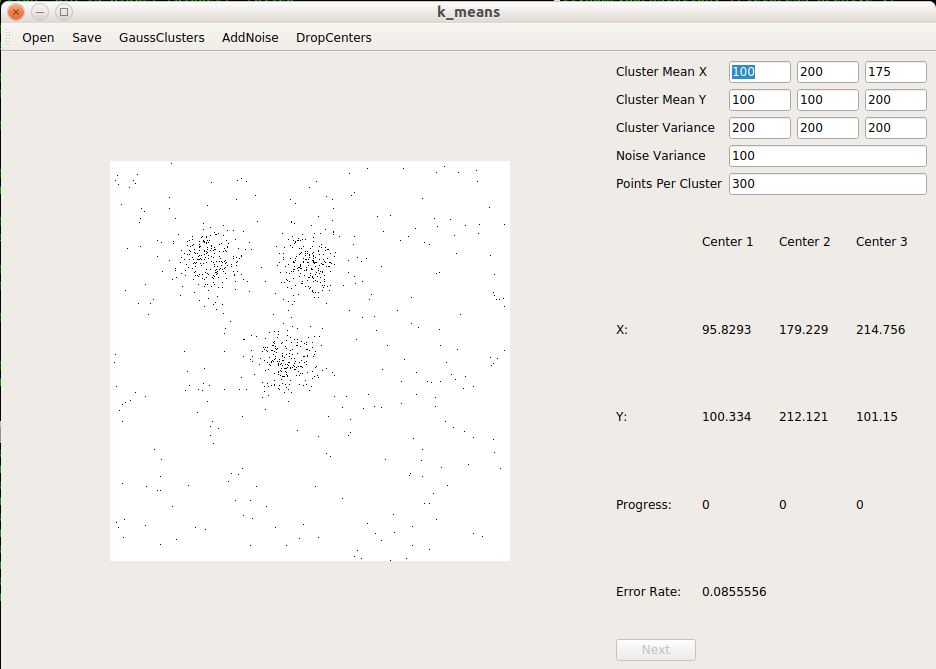
## 实验目的

1. 使用K-means算法对随机生成的高斯点集聚类。
2. 使用K-means算法对叠加了噪声的高斯点集聚类。
3. 观察噪声对聚类误差率的影响。

## 实验过程

1. 设置好三个高斯点集的各种参数，随机生成三个高斯点集。尝试生成不同期望、方差、点数的高斯点集合。观察形态。
2. 随机投入三个聚类中心的初始位置，单步运行，观察动态收敛过程。尝试多次随机投入中心，并单步运行至收敛，观察收敛过程和效果。并记录聚类误差率。
3. 设置噪声强度（方差），添加进已有点集。尝试不同的噪声强度，观察形态。
4. 对上述已经添加噪声的进行与第2步相同的操作。

## 实验现象

1. 高斯点集的一组参数取值和生成形态。  
   
2. 随机投入聚类中心点后，单步运行并逐渐收敛的过程。  
   
3. 对上述高斯点集叠加的噪声强度和叠加噪声以后的点集形态。  
   
4. 再次随机投入聚类中心，单步运行并逐渐收敛。基本流程与不加噪声相似。

## 实验结果

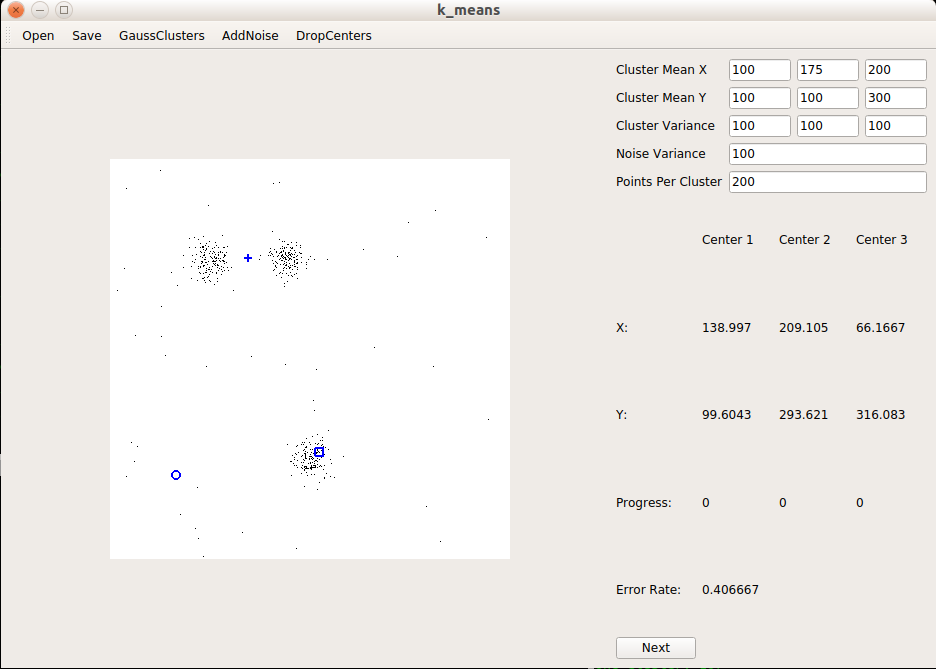
每个点集300个点，期望分别位于（100, 100），（200, 100），（175, 200）时，改变方差，在相同方差的情况下多次随机生成。方差与聚类错误率的关系：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 第1组 | 第2组 | 第3组 | 第4组 | 第5组 | 第6组 | 平均 |
| 100 | 0.1633 | 0.1488 | 0.1588 | 0.1711 | 0.1533 | 0.1166 | 0.15198 |
| 200 | 0.0755 | 0.0722 | 0.0877 | 0.0888 | 0.0677 | 0.0800 | 0.07865 |
| 300 | 0.0722 | 0.0611 | 0.0777 | 0.0544 | 0.0677 | 0.0466 | 0.06328 |
| 400 | 0.0477 | 0.0711 | 0.0488 | 0.0544 | 0.0500 | 0.0455 | 0.05291 |
| 500 | 0.0577 | 0.0533 | 0.0411 | 0.0600 | 0.0622 | 0.0488 | 0.05385 |
| 600 | 0.0566 | 0.0466 | 0.0688 | 0.0611 | 0.0688 | 0.0455 | 0.05790 |
| 700 | 0.0544 | 0.0700 | 0.0622 | 0.0644 | 0.0688 | 0.0400 | 0.05997 |
| 800 | 0.0744 | 0.0811 | 0.0666 | 0.0766 | 0.0722 | 0.0688 | 0.07328 |
| 900 | 0.0811 | 0.0755 | 0.0855 | 0.0811 | 0.0833 | 0.0655 | 0.07867 |
| 1000 | 0.1122 | 0.1088 | 0.0911 | 0.1055 | 0.1033 | 0.0966 | 0.10292 |

每个点集300个点，期望分别位于（100, 100），（300, 100），（175, 250），方差均为200时，生成三个原始高斯点集，对这三个原始的高斯点集加不同强度的噪声。噪声强度与聚类错误率的关系：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 第1组 | 第2组 | 第3组 | 第4组 | 第5组 | 第6组 | 平均 |
| 0 | 0.0766 | -- | -- | -- | -- | -- | 0.07660 |
| 200 | 0.1666 | 0.1522 | 0.1566 | 0.1466 | 0.1455 | 0.1377 | 0.15086 |
| 400 | 0.1066 | 0.1077 | 0.1188 | 0.1088 | 0.1100 | 0.1077 | 0.10993 |
| 600 | 0.1033 | 0.0922 | 0.0911 | 0.0977 | 0.1044 | 0.1066 | 0.09921 |
| 800 | 0.0933 | 0.0900 | 0.0966 | 0.0988 | 0.0877 | 0.1000 | 0.09440 |
| 1000 | 0.0955 | 0.0944 | 0.1022 | 0.0877 | 0.0977 | 0.1033 | 0.09680 |
| 1200 | 0.1066 | 0.0911 | 0.0966 | 0.1022 | 0.0977 | 0.1066 | 0.10013 |
| 1400 | 0.1055 | 0.1066 | 0.1022 | 0.1000 | 0.1077 | 0.1111 | 0.10551 |
| 1600 | 0.1155 | 0.1077 | 0.1122 | 0.1222 | 0.1200 | 0.1044 | 0.11367 |
| 1800 | 0.1144 | 0.1133 | 0.1255 | 0.1244 | 0.1066 | 0.1100 | 0.11570 |
| 2000 | 0.1288 | 0.1300 | 0.1277 | 0.1266 | 0.1211 | 0.1244 | 0.12643 |

## 遇到的问题

以上实验数据都是在目测已经获得全局最优解的时候记录的，如果聚类中心收敛后，看起来不像全局最优解，那么认为是聚类中心初始投入点不够好，本组作废，进行下一组。局部最优解，除了其聚类中心与设定的高斯点集期望相差较大这个特点以外，另一个明显的特征就是聚类错误率过高。

上图所示的情况，很明显是局部最优解。聚类中心（右下）与高斯点集期望（右上）相差极大。而且聚类误差率达到了40%以上。

## 一些猜测

在完成上述在其他条件相同的情况下，若使用欧几里得距离做聚类，更易收敛至局部最优解的情况：

1. 不同高斯点集的期望分布密集。
2. 单个高斯点集方差较小。
3. 单个点集点数少。
4. 一个或多个聚类中心初始位置远离三个点集所有点的重心。
5. 加入噪声强度小或没有噪声。