**报告**

**学号: 姓名: 班级:**

1. **目标**

练习如何进行聚类、聚类分析和聚类可视化。熟悉开发系统、其图形用户界面和输入数据格式。

1. **数据**

数据集是[GSE7390\_transbig2006affy\_demo.csv](http://web.cs.wpi.edu/~ruiz/Courses/bcb4003_503/AF11/Projects/Project1/GSE7390Dataset/GSE7390DemographicData/GSE7390_transbig2006affy_demo.txt)。其中包含了TRANSBIG验证研究中198名未治疗患者的信息。请参阅介绍数据集的 README.txt 文件。

1. **实验**

**任务 1**

1. **预处理**
2. 认识你的数据。数据中有多少个实例、名义属性、数字属性？也就是要熟悉数据集的不同属性，它们的分布情况。用表格列出数值属性的范围。是否有一些数值属性没有用？如果有，就删除这个属性
3. 有些名义属性有太多不同的值。它们不包含有用的信息。将它们从考虑中删除，以创建一个新的.arff数据集文件。列出被删除属性的名称。
4. 规范化数据，创建一个新的.arff数据集文件。在你的.arff文件中列出数据集的前10个数据实例。
5. **K-means**
6. 设定不同的k: k = 2 to 6,
7. 使用不同的距离指标: Euclidean and Manhattan (= cityblock).
8. 使用不同的种子值: 10, 27, 43;
9. 在聚类前对每个属性进行标准化和不进行标准化.

使用一个表格来形成每个实验及其结果的摘要：使用的参数，每个聚类中的实例数，中心点，以及每个聚类的误差值（"聚类内平方误差之和"）。你对每个聚类有什么有趣的发现或观察。

1. **聚类分析**

选择最佳聚类（即误差值最低）。制作结果的散点图可视化。

**任务2**

Practice to code for k-means algorithm with the following dataset. The Number of clusters is 3. The distance function is [Euclidean](C:/Program%20Files/YoudaoDict/6.3.69.8341/resultui/frame/javascript:void(0);) [distance](C:/Program%20Files/YoudaoDict/6.3.69.8341/resultui/frame/javascript:void(0);). List the initial data, the center of the iteration 1, and the final clusters.

|  |  |  |
| --- | --- | --- |
| ID | Feature 1 | Feature 2 |
| 1 | 2 | 10 |
| 2 | 2 | 5 |
| 3 | 8 | 4 |
| 4 | 5 | 8 |
| 5 | 7 | 5 |
| 6 | 6 | 4 |
| 7 | 1 | 2 |
| 8 | 4 | 9 |