1、问题描述

给一份数据源、已知该数据源中有三类数据，并且每一类数据都已经标明类编号。现在要求在抹去数据源中类编号的情况下，使用 K-means 及 K 中心算法对原数据进行聚类，得到每一个数据的类编号，并在最后与源数据集进行比对得出聚类的准确率。

2、数据源描述

名称：葡萄酒识别数据

来源：Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.

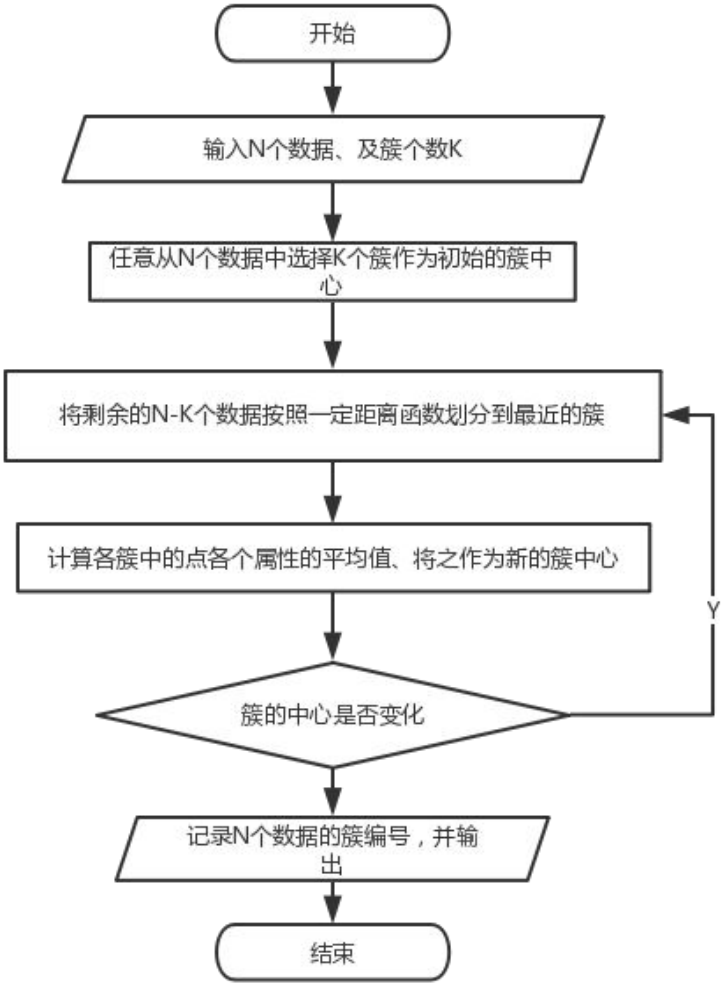
实例数量：178

属性数量：13

每类实例：1：59；2：71；3：48

3、K-means 算法

1)算法流程图



2)算法描述

输入：n 个数据的数据集合和已知的簇个数 k

输出：n 个数据各属于 k 个簇中哪个簇的信息

算法步骤：

1）任意从 n 个数据中选择 k 个作为初始的簇中心；

2）将剩余的 n-k 个数据按照一定的距离函数划分到最近的簇；

3）repeat

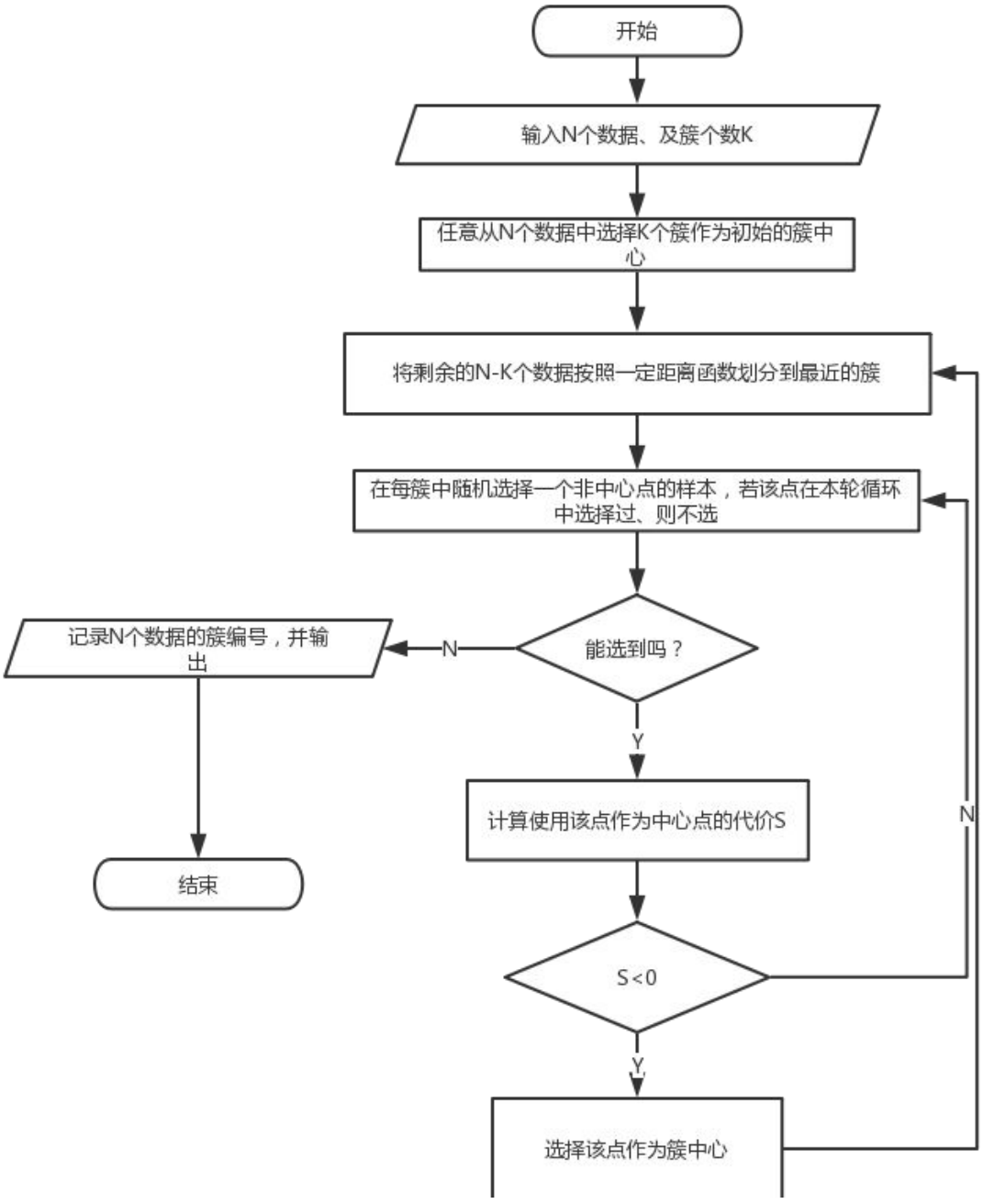
4）按一定的距离函数计算各个簇中数据的各属性平均值，作为新的簇中心；

5）重新将 n 个数据按照一定的距离函数划分到最近的簇；

6）until 簇的中心不再变化。

4、K 中心算法

1)算法流程图



2)算法描述

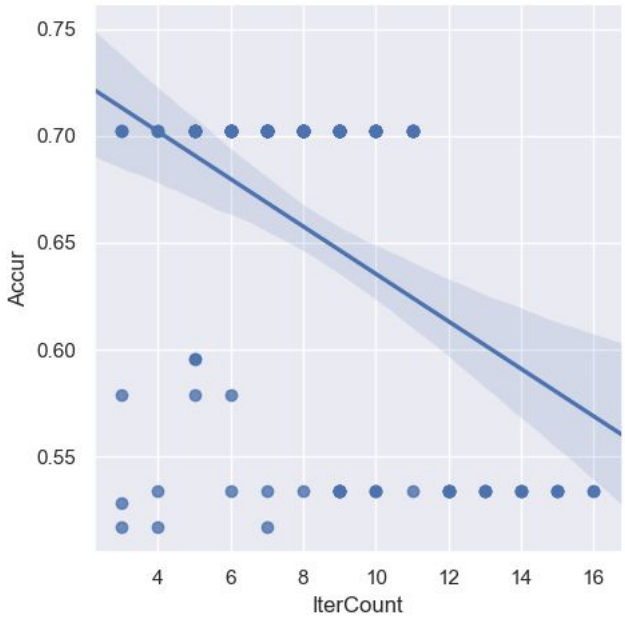
输入：簇的个数 k，包含 n 个样本的数据集

输出：各样本属于 k 个簇的信息

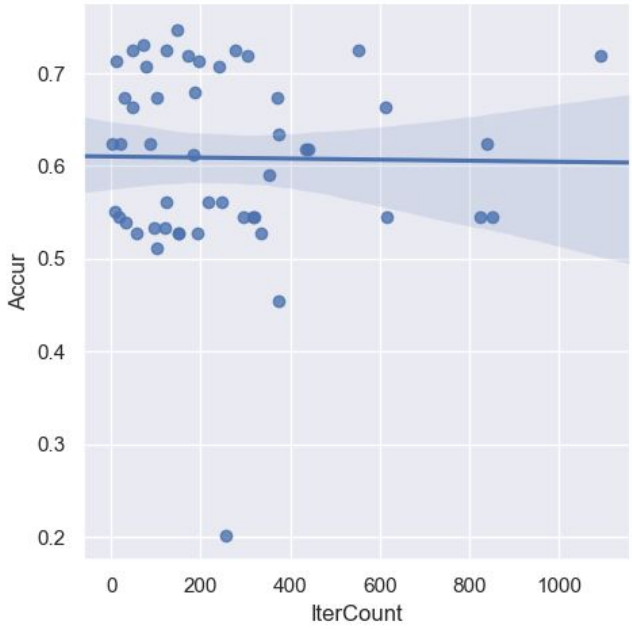
算法步骤：

1）随机选择 k 个样本作为初始中心点；
2）repeat
3）将非中心点的数据依照与各中心点的距离划分到最近的簇中；
4）随机的在非中心点中选择一个样本；
5）计算使用该点做中心点来代替原中心点的代价；
6）if <0 then 用该点替换原中心点，形成新的簇集合 else 继续寻找
7）if all >0 then 不改变中心点
8）until 中心点不再发生变化

5、各算法的 Implot 图



K-means 算法统计图



K 中心算法统计图

6、实验结果分析

实验结果及其不理想、显然需要改进、我添加了为数据进行预处理的过程，利用 Min-Max Scaling 进行归一化、但结果依旧不理想，后续慢慢改进。