k-means 聚类算法在吸收工段中的应用

1 问题背景

化工行业具有生产过程复杂、生产工艺多样以及生产安全极端重要的特点，作为化工专业的高素质人才，需要的不仅仅是化工的专业知识，工程经验的积累也是不可忽视的一个方面。但在实际的工作过程中，操作人员可能因为种种原因离职，带走大量的操作经验，给企业带来损失。随着的工业现场总线技术的发展和计算机集散控制系统日渐成熟，使得大量的工业数据可以通过成熟的数据采集和存储系统保存下来。在此基础下，利用数据挖掘技术能够从过程数据中发现操作人员的操作经验并作为知识保存，在一定程度上减少了损失。

2 问题描述

在不同的工况下，目标变量被控制在不同的状态，如果对工况不了解，就无法对目标变量进行分类。例如，在图 1 所示的吸收过程中，吸收剂 A 和气体 D 的进料量会对吸收状态有着较大的影响，在对吸收的具体情况不熟悉的条件下，利用吸收剂 A 与气体 D 的实时数据，可以用聚类算法对这两个变量进行聚类，从而初步衡量操作情况。在聚类的结果中，每一个聚类中心反映了一类运行状态。

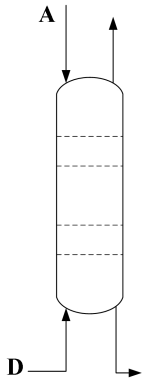

图 1 吸收过程示意图

3 方法与结果

首先读取吸收剂 A 与气体 D 的进料流量数据，共计 500 条。利用 Python 中的 matplotlib 绘图工具库以吸收剂 A 的进料量为 x 轴，气体 D 的进料量为 y 轴绘图查看数据的分布情况，如图二所示：

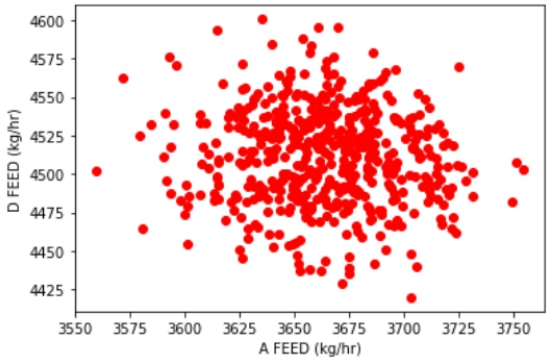

图 2 数据分布情况

从图中并不能直观的看出数据类别，这里使用 Python 中的 sklearn 工具库，调用其中的 k-means 算法来进行聚类。首先尝试聚类为 2 簇的情况，如图 3 所示，数据被划分成了两个类别，其中红色五星代表了聚类中心。

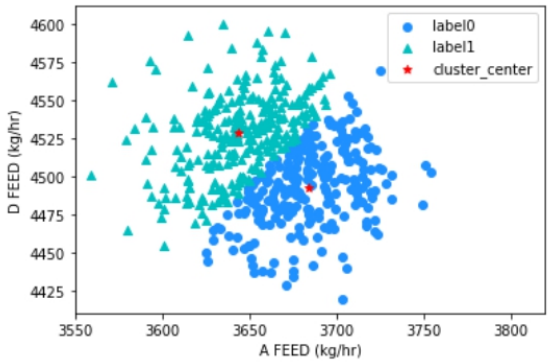

图 3 n_clusters=2(2 簇)的聚类情况

将聚类的簇数量设置为 3 时，算法依旧能够给出不错的结果（图 4）。

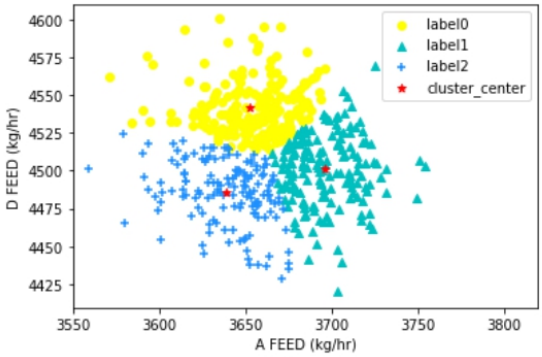

图 4 n_clusters=3(3 簇)的聚类情况

从上面的例子可以看出，k-means 聚类算法虽然能够按照我们的要求完成数据聚类的任务，然而簇数量的确定依然需要根据实际的生产状况来确定。所以说，数据挖掘技术不能脱离化工生产实际，两者的结合才能创造更多的价值。