weka聚类

201250203 陈张熠

一. 使用方法理解

K-Means算法是一种迭代型聚类算法，采用欧式距离作为相似性指标，发现给定数据集中的K个类，且每个类的中心是根据类中所有数值的均值得到的，每个类的中心用聚类中心来描述。对于一个给定的数据集，随机选取K个样本作为中心，计算各样本与各个聚类中心的距离，将各样本回归于与之距离最近的聚类中心，求各个类的样本的均值，作为新的聚类中心，判定若类中心不再发生变动或者达到迭代次数，算法结束，继续计算距离。

本实验中主要使用了weka中六种方法：

1. simpleKMeans.setNumClusters(3)：设置k-means的群数为5，可以将数据集中的数据分为3类
2. simpleKMeans.buildClusterer(data)：使用simpleKMeans进行分析数据
3. simpleKMeans.setMaxIterations(5)：设置simpleKMeans最大迭代数量

二. 数据集处理的思路

通过创建FileReader对象，并让其读取对应的arff文件。

通过此Reader对象，创建Instances对象

此后，simpleKMeans.setNumClusters(3)为数据集设置群数，可以将数据集分为3类，用simpleKMeans.setMaxIterations(5)方法设置最大迭代数为5

最后通过通过SimpleKMeans中的buildCluster进行分析数据

三. 实验结果