**聚类学习报告**

**姓名：闫庆菲**

**学号：320180942321**

**摘要：聚类是一种典型的无监督学习任务，它将数据集中相似的样本分到相同的簇中，每个簇中的样本对应一个潜在的类别。本次任务中主要使用了K-Means、Birch、高斯混合模型这三种聚类模型。**

**1.引言：**本次工作主要分为两个部分，第一部分是在鸢尾花数据集【1】、乳腺癌数据集【2】、预处理后的鲍鱼年龄预测数据集【3】这三个数据集上使用自主实现的K-Means算法、高斯混合模型以及Birch算法进行聚类分析；第二部分是实现RI和purity两个评价指标的计算，对聚类结果进行评价。

**2.算法：**

**2.1.K-Means：**

**K-Means是一种应用比较广泛的聚类分析方法，它的目标是将n个样本划分到K个簇中。因此，首先要从数据集中随机选择k个不同的数据点作为初始的簇中心，然后计算每一个数据点到每一个簇中心的距离，将其归入距离最小的那个簇，这样遍历完一遍数据集后将形成k个簇，然后计算每个簇的均值作为新的簇中心，重复已上操作，直到簇中心不再发生变化时停止。**

**2.2.高斯混合模型：**

高斯混合模型假设数据集D是由多个服从不同期望和标准差的正态分布混合在一起的，不过每个分布的参数、比例都未知。每个分布都对应了一个簇，因此只要根据数据集D确定了每个分布的参数，就可以确定每个样本属于哪个簇。而求解参数需要用到EM算法，它用迭代的方式求解，首先随机初始化一个模型参数，然后建立对数似然函数的下界为的期望，取得下界时，对每一个样本，计算隐变量的概率分布，之后最大化对数似然函数的下界获得此时的：

重复以上步骤直到收敛。

**2.3.Birch：**

Birch是利用层次方法的平衡迭代规约和聚类方法，它利用了聚类特征树CF Tree来帮助我们快速实现聚类，CF Tree的每一个节点由若干个聚类特征CF组成，聚类特征为一个三元组，其第一个元素为该CF中包含的样本量，第二个元素为该CF中拥有的样本点各特征维度的和向量，第三个元素为该CF中拥有的样本点各特征维度的平方和。对于CF Tree，我们一般有几个重要参数，第一个参数是每个内部节点的最大CF数B，第二个参数是每个叶子节点的最大CF数L，第三个参数是针对叶子节点中某个CF中的样本点来说的，它是叶节点每个CF的最大样本半径阈值T，也就是说，在这个CF中的所有样本点一定要在半径小于T的一个超球体内。而当叶子节点的CF数超出规定的最大CF数时，将叶子节点分裂即可。

**3.实验及结果分析：**

为了对实验结果进行评价，我实现了两个方法分别计算聚类结果的纯度和兰德系数，之后将得到的结果绘制成了图3-1、图3-2以及图3-3。图3-1是鸢尾花数据集上三种聚类方法的指标可视化结果，从图上我们可以看出，不论是纯度还是兰德系数，高斯混合模型的指标值都要高于其它两个模型，其次是K-Means模型，最后是Birch模型，所以在鸢尾花数据集上高斯混合模型的聚类效果最佳。图3-2是癌症预测数据集上三种聚类方法的指标可视化结果，从图上我们同样可以看出，不论是纯度还是兰德系数，高斯混合模型的指标值都要高于其它两个模型，其次是K-Means模型，最后是Birch模型，所以在癌症预测数据集上高斯混合模型的聚类效果最佳。图3-3是鲍鱼年龄预测数据集上三种聚类方法的指标可视化结果，从图上我们同样可以看出，不论是纯度还是兰德系数，高斯混合模型的指标值都要明显高于其它两个模型，而其它两个模型的指标相差不大，不过Birch的值要略高于K-Means，所以在癌症预测数据集上高斯混合模型的聚类效果最佳。

**4.结论：**

本次工作让我对聚类算法有了一个更加深刻的认识，经过自己实现K-Means，加深了我对该算法的理解，通过三种算法在同一数据集上聚类结果的对比分析，我对三种算法的效果也有了新的认识。

**参考文献：**

**【1】**[Index of /ml/machine-learning-databases/iris (uci.edu)](http://archive.ics.uci.edu/ml/machine-learning-databases/iris/)

**【2】**[Index of /ml/machine-learning-databases/breast-cancer-wisconsin (uci.edu)](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/)

**【3】**[Index of /ml/machine-learning-databases/abalone (uci.edu)](http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/)