聚类

陈晓宇 MF1633007 traumcxy@163.com

1. 实现细节

通过k-Medoids聚类,spectral聚类实现对已有数据集的聚类，下面分别给出这两种聚类的伪代码。

（1）k-Medoids聚类：

创建k个点作为起始质心

当任意一个点的簇分配结果发生改变时

对数据集中的每个数据点

对每个质心

计算质心与数据点之间的距离

将数据点分配到距离其最近的簇

对每一个簇，计算簇中到各个点的距离和最小的点作为质心

（2）谱聚类：

初始化：邻接矩阵W，距离矩阵Distance初始化为全0矩阵，k为聚类的个数，m为数据点的个数

预处理：计算出数据集中任意俩个点之间的距离，构造出距离矩阵Distance

对于数据集中的每一个点xi

找出与数据点i距离最近的n个点

对于n个点中的每一个点j

将距离矩阵更新，即：W[i, j] = W[j, i] = 1

计算D矩阵和拉普拉斯矩阵L：D是一个对角阵，第i个元素为W的第i列的所有元素之和；L=D-W

计算矩阵X=inv( D ) \*L，计算出矩阵X的所有特征值value与特征向量vector

对于vector，按照其对应的特征值从小到大进行排序，得到y0,y1,y2…..ym-2,ym-1

选取特征向量y1,y2,…,yk，组成降维后的数据集dataSet = [ y1 y2 … yk ]

对于降维后的新的数据集，在其上使用k-Medoids算法进行聚类。

1. 结果
   1. 下为利用purity得到的结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| purity | k-means | Spectral(n=3) | Spectral(n=6) | Spectral(n=9) |
| Dataset 1 | 0.7 | 0.699 | 0.699 | 0.699 |
| Dataset 2 | 0.597 | 0.624 | 0.591 | 0.592 |

* 1. 下为利用gini得到的结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| gini | k-means | Spectral(n=3) | Spectral(n=6) | Spectral(n=9) |
| Dataset 1 | 0.412 | 0.419 | 0.419 | 0.419 |
| Dataset 2 | 0.542 | 0.431 | 0.454 | 0.462 |