题目: Clustering

胡求 MG1633031 huqiu00@163.com 188-5182-3759

(南京大学 计算机科学与技术系, 南京 210023)

1 实现细节

1. 本文描述实现两种聚类算法，其一为 K-medoids （K-中心点）算法，其二为基于谱图理论的谱聚类算法。数据集有两个，一个是小数据集 german.txt，其中有 1000 个样本，每个样本有 24 个特征，每个样本占一行，最后为该样本的标签，标签分两类，1 和-1，标签用作 Gini 系数和 Purity 的计算。另一个是大数据集 mnist.txt，其中包括 10000 个样本，每个样本有 784 个特征，每个样本占一行，最后为该样本的标签，标签分 10 类，0-9。下面简要说明两种聚类算法的流程。
2. K-medoids 算法：类似 K-means 算法的原理，只是聚类的中心不是所有该类所有样本的特征平均值，而是限制聚类的中心必须是该类的某一个样本，这样做的好处是可以有效避免离群点的影响，适用于异构数据等。初始找 k 个中心点，然后将每个点分到离它最近的中心点代表的类中，然后对每一类更新中心点，这样循环往复直到中心点不再变化。

   算法伪代码如下：

输入： 数据矩阵 dataSet
聚类个数 K

过程：

```
随机选取 K 个不同的样本作为聚类中心点,{m1,m2...mK}
while True：
    for i in 1...N do
        将每个样本归到其最近中心点所代表的类
    end for
    for k in 1...K do
        选取第 k 类中的一个点 i，使这个点到其他该类中的点的距离和最小
        将第 k 类的中心点替换为 i
    end for
    if 所有类中心点无变化：
        break
```

输出： 最终聚类中心点$\{m_1^*, m_2^*, \ldots, m_k^*\}$

---

3. 谱聚类算法，基于谱图理论，适用于数据是非凸的或者类似于嵌入高维空间的低维流形时。基本思路是对每个样本点，与跟它最相似的多个样本点之间连边，边的权值可以赋为相似度或者 1（两种方式），然后计算拉普拉斯矩阵 L=D-W，其中 D 为每个样本点的度

作为对角元素组成的对角矩阵，W 就是刚才的距离矩阵。然后对拉普拉斯矩阵 L 进行特征分解，取最小的 k 个特征值对应的特征向量作为新的样本表示，然后对这个样本表示进行 K-medoids 聚类或者 K-means 聚类等即可得到聚类结果。
算法伪代码如下：

输入： 数据矩阵 dataSet
聚类个数 K
近邻连接图近邻数 n

过程：

1： 对每个数据点，计算 n 近邻的点集$\{p_1, p_2, \ldots, p_n\}$

2： 将每个数据点与其 n 个近邻各连一条无向边 e
3： **for** every edge (u, v):
4： W[u][v] = W[v][u] = 1
5： **end for**
6： sigma_W = sum(W, axis=0) ,对 W 所有列求和得 sigma_W
7： 构造对角矩阵 D = diag(sigma_W)
8： 构造拉普拉斯矩阵 L=D - W
9： 对 L 进行特征分解， eval, evec = eig(L)
10：选取最小的 K 个特征值对应的特征向量，组成新的数据矩阵 new_datamat
11：K-medoids(new_datamat, k)

输出： 聚类中心点$\{m_1^*, m_2^*, \ldots, m_k^*\}$

2 结果

2.1 实验设置

1. 数据来源为从作业网站提供的 german 和 mnist 数据集，german 是小数据集，mnist 是较大的数据集，数据说明见上一节。
2. 采用 Python 2.7.12 和 Ubuntu16.04 系统 作为编程环境
3. 输出的结果将会被组成一个表格在下文给出。
4. 结果数据包括在不同数据集下 k-medoids 算法聚类的效果，以 Gini 系数和 Purity 作为衡量依据，以及在不同数据集下和不同近邻数下 Spectral Clustering 聚类的效果，结果也以 Gini 系数和 Purity 作为依据。

2.2 实验结果

测试结果如下，不同数据集下，不同算法下进行聚类的 Purity 见图 1。

Purity	k-medoids	Spectral(n=3)	Spectral(n=6)	Spectral(n=9)
german 数据集	0.7	0.7	0.7	0.7
mnist 数据集	0.4949	0.7385	0.7253	0.714

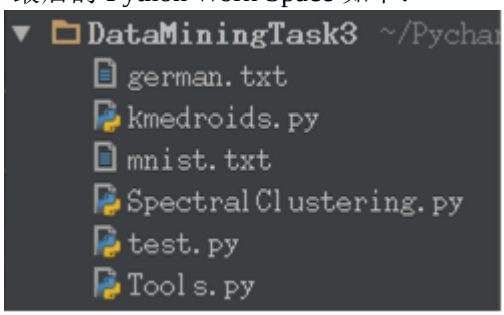
图 1 Purity 评测

不同数据集下，不同算法下进行聚类的 Gini 系数见图 2。

Purity	k-medoids	Spectral(n=3)	Spectral(n=6)	Spectral(n=9)
german 数据集	0.3903	0.4158	0.4132	0.4159
mnist 数据集	0.6508	0.3436	0.3569	0.368

图 2 Gini 评测

最后的 Python Work Space 如下：



其中 getman.txt, mnist.txt 为数据集文件。
kmedoids.py 为 K-medoids 算法的实现和测试。
SpectralClustering.py 为谱聚类算法的实现与测试。
Tools.py 为一些全局的工具函数。
test.py 为一些小测试，内容与程序无关。

（实验结果后可简述对当前实验的思考）

2.3 思考

1. 纯枚举中心点替换的方法虽然说准确度高，但是速度太慢，在更大的数据集下必然很低效，可以考虑采用随机选取 r 个点对进行替换的方式，损失一点准确度，来对算法进行加速。或者也可以采用其他优化算法优化效率。

2016.10.31