关于论文《Robust continuous clustering》的学习报告

当前存在的多数聚类算法都是 center-base 的算法，这些算法普遍对初始化的方式较为敏感、不适用于高维空间中、需要预先定义聚类的个数等缺点。为了克服这些缺点，作者们在这篇论文中提出一个新算法，这个新算法高效、能扩展到高维的数据中，并且不需要预先定义聚类的个数，应用到不同的数据集中都取得了很好的效果。

论文把聚类的过程变成一个基于抗差估计的连续目标函数的优化过程，尽管这个目标函数是非凸的，作者们在优化求解的过程中仍然使用了线性最小二乘求解方法，这一点是应用得比较巧妙的求解方式。

论文提出的算法的大概流程是对于一个特定的数据集 X，目标是找到这个数据集中每个数据点的代表 U。优化的过程是迭代式地把多个数据点x_i的表示u_i就可能地合并成一个相同的表示，这个相同的表示就是这些数据点的代表点，而这些具有相同代表的数据点就被聚为同一类。这一算法克服了 center-base 算法的预先定义聚类个数的缺点，这是因为最终的代表点的个数就是最终的聚类个数。

RCC 算法产生了以下的目标函数：

$$\mathbf{C}(\mathbf{U}) = \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \frac{\lambda}{2}\sum_{(p,q)\in\mathcal{E}} w_{p,q}\,\rho\big(\|\mathbf{u}_p - \mathbf{u}_q\|_2\big). \quad [1]$$

前面提到我们最终的目标是把所有数据点 X 表示成 U，从而使得具有相同代表u_i的数据点属于同一类。从目标函数上解释这个目的：目标函数的第一项惩罚的是每个数据点与该数据点的表示之间的差距，这一点不难理解，每个点应该以某个距离自己不远处的点作为代表；而目标函数的第二项是这个算法的关键，首先解释以下变量ε，这是一个边集，但是对于一个给定的数据集，并不是一个图，因此数据集本身不存在边，论文中采用的方法是构建一个 m-KNN 图（关于 m-KNN 图的资料比较少，因此我理解成基本的 KNN 图），构建的具体方法是，计算每个数据点v_i与数据集中其他数据点之间的数据，选取距离最近的 K 个数据点，为数据点v_i与这 K 个数据点之间增加一条边，这就形成了基于数据集的 KNN 图。目标函数的第二项惩罚的是相隔较近数据点的代表是否一致，直观上分析，相隔较近的数据点很大可能会被划分为同一个簇，因而相隔较近的数据点的代表应该尽可能地一致。至于为什么要构建 KNN 图而不是为每对数据点都增加一条边？我的理解是每个数据点只会以附近不远的点作为代表点，而相邻较远的点的代表点肯定不一样，因此这一衡量没意义。

对 RCC 算法的目标函数的求解过程的部分理解：增加了一个辅助变量 L 把原始 RCC-Objective 转化成新的目标函数 C(U,L)，然后再利用目标函数的 biconvex 性质，固定一个变量从而单独求解另一个变量。而这里的 biconvex 性质是通过初始设定μ值得到的，在迭代的过程中，μ值逐渐自动减少，慢慢地把非凸的性质引入到目标函数中，最终求到一个与最优值很接近的解。（虽然不了解这样做的原理，但是觉得这种思想本身就很巧妙）

RCC-DR 算法：前面学习的 RCC 算法利用的是把数据集的每个数据点表示成这个数据点的代表的形式，而数据点的代表与原来的数据点的维度是一样的。在论文的后半部分，作者们把原始 RCC 算法与降维技术结合在一起，把数据点的代表表示到一个低维空间中，最后在低维空间中进行聚类。RCC-DR 的目标函数如下：

$$\mathbf{C}(\mathbf{U},\mathbf{Z},\mathbf{D}) = \|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_2^2 + \gamma\sum_{i=1}\|\mathbf{z}_i\|_1 \qquad [10]$$

$$+\nu\left(\sum_{i=1}^{n}\|\mathbf{z}_i - \mathbf{u}_i\|_2^2 + \frac{\lambda}{2}\sum_{(p,q)\in\mathcal{E}} w_{p,q}\,\rho\left(\|\mathbf{u}_p - \mathbf{u}_q\|_2\right)\right).$$

首先我们对原始数据集的所有数据点进行稀疏编码，将 X 降维表示成 Z。RCC-DR 目标函数的第一项惩罚的是稀疏编码的过程中为了使 Z 尽可能多地包含原来数据集 X 的信息。而第二项我认为是一个正则化项，是为了防止过拟合。而第三第四项与原始 RCC 算法的目标函数是惩罚作用是一致的。