关于论文《Automatic Clustering via Outward Statistical Testing on Density Metrics》的学习报告

1. 关于 RLClu 算法：该算法基于两个参数表征一个数据点作为聚类中心的可能性：第一个是局部密度，表征了该数据点附近点的密集程度，可是局部密度不能唯一表征一个聚类中心，因为可能存在两个具有高局部密度的点距离很近，这样也只是属于同一个类。于是，算法的作者们很巧妙地提出了第二个参数——高密度最近零距离，表征了局部密度比某个数据点大的点中离它最近的那一个点。考虑这样一种情况，如果一个点拥有较大的高密度最近临距离，表示密度比它大的点都离它较远，如果此时该点的局部密度也较大（排除掉离散点的情况），那么这个点作为聚类中心的可能性就很大。两个参数的共同作用就正是这个算法的微妙之处，这个算法思想真是让人叹为观止。
2. 算法的缺陷 :首先是局部密度的计算 ,原算法提出的计算方法是统计以某数据点为圆形，固定长为半径的圆内有多少数据点，我觉得这种计算方法是有一定的局限性的，因为它统计的是一定区域范围内的数据点数，而忽略了距离对局部密度产生的影响，很容易想象得到，距离越近，对局部密度的贡献就越大，因此原算法的不基于距离的密度计算方法是有待提高的。再者就是参数 dc 的选择，由论文的研究可以看出算法的计算方法对这个参数是很敏感的。因此，论文提出了一个新的局部密度计算方法，该方法体现了距离对密度的作业同时对 dc 的选择具有很大的健壮性。而除了该计算方法以外，我还想到了高斯核函数，高斯核函数服从正态分布，距离越近函数值越大，很重要的一个特性是当自变量大于截断值 dc 时，函数衰减得很严重，因此对于离得较远的数据点，对局部密度的贡献值就越低，但是也会有一点小影响，我觉得这样是更符合实际情况的，当然，这只是我的一个想法因此算法对于 dc 的健壮性还不清楚。再者就是需要预先设定聚类中心的个数，这一点后面再讨论。
3. 从 RLClu 算法中我们知道，聚类中心的特性就是同时拥有较大的局部密度和高密度最近临距离的数据点，因此，在论文中，作者们提出的 STClu 算法用这两个参数的乘积γ对聚类中心进行表征。而由于聚类中心的高密度最近临距离σ相对与其他数据点会更大些，而且聚类中心的数目明显少于其他数据点，因此 STClu 算法的思想是找到那些相对于其他数据点的γ偏离的点（论文中成为 outliers），这就是最终要寻找的聚类中心。
4. 论文是采用的是统计学和概率论的方法来寻找这些点。首先，由于高密度最近临距离σ是服从长尾分布的(这里的分析证明我也没看懂)以及局部密度大于 0 的概率为整数，根据定理 1，可以证明γ是服从长尾分布的。接下来是运用长尾分布特性在长尾分布的数据中找出其中的 outliers，也就是聚类中心。
5. 算法的流程如下：首先就γ按照降序排列，形成数组 X；接着，引入一个变量$H_{0,k}$，表示数组 X 的第 k 个元素$X_{1,k}$，属于长尾分布的猜想。算法的原理是，假如这个猜想被 rejected，那么在该长尾分布中，这个$H_{0,k}$对应的数据点以及所有下标比 k 小的$H_{0,k}$对应数据点都属于聚类中心。而判断一个猜想是否被 rejected 的标准是运用公式$R_m > r_m$是否成立。

   感想：这个算法实现起来好像不太难，可是算法的思想太巧妙和数学逻辑太强了。首先

是原算法中用于表征聚类中心的两个参数，直接用这两个参数能避免多次迭代重复寻找聚类中心的过程。其次就是运用统计学和概率论的知识对新算法进行证明与应用，先是证明了γ是服从长尾分布的，这种操作很数学，这样就把原来的寻找聚类中心的算法转化到一堆服从长尾分布的点中寻找其中的 outliers，这些点就是最终的聚类中心。看完这篇论文，对这个算法是有一定的了解，以及明白了算法的大致流程，但是受限于数学知识，对算法的原理以及推导还没能很好地理解。还有一点就是下面这个函数，看完了整个算法还是不太了解这一部分是用来做什么以及这个函数的作用。

In order to identify the outliers, we need to find an effective statistic to test the sequence of hypotheses $\{\mathrm{H}_{0,k}, 1 \leq k \leq m\}$. Considering that i) the power function can be used to describe the distribution of the product of local density ρ and the minimum density-based distance δ under a proper setting of cutoff distance d_c [1], and ii) Rodriguez and Laio also stated that in the region with both large ρ and δ, the distribution will be strikingly different from the power law, and the high values of the product would be more likely to be outliers [1]. Meanwhile, comparing to ρ and δ, the new metrics $\hat{\rho}$ and $\hat{\delta}$ proposed in this paper enhanced them in robustness while still follows the idea in [1] that the objects with both of larger ρ and δ are more possible to be the clustering centers. So, it is rational to assume that the tails of the long-tailed distribution (e.g., $\hat{\gamma} = \hat{\rho} \times \hat{\delta}$) which decay as power functions. Specifically, the cumulative density function of F can be defined as follows, for some $\lambda > 0$ and sufficiently large x,

$$F(x) = 1 - L_0(x) \cdot x^{-\lambda}, \tag{6}$$

where L_0 is a slowly varying function and for sufficiently large x, L_0 behaves almost like a constant, and the parameter λ denotes the tail index.