关于论文《LWMC: A Locally Weighted Meta-Clustering Algorithm for Ensemble Clustering》的学习报告

论文主要解决的问题是进行聚类集成，即把多个聚类成员根据其相似度进行加权，从而把这些聚类成员的信息进行整体考虑，得到一个效果更优，更具鲁棒性的整体聚类结果。

算法的具体流程如下：对于给定的数据集 N 个点，首先可以使用不同的聚类算法或同一个聚类算法的不同参数设置生成多个聚类结果，成为基聚类**π**，接着把所有基聚类中的所有聚类生成一个簇集合 C。然后，两个簇之间的相似度用 jaccard similarity 来表征，接着，我们把簇作为点、两个簇间的相似度作为边的距离构建一个 CSG 图，再通过 Ncut 算法对 CSG 图进行切割，生成的每一部分就是一个超聚类，也就是整合聚类的结果。最后要进行的工作就是把这 N 个数据点按照相应的规则归类到这些超聚类中。论文中采取的方法是采用香农熵的基本规则计算把每一个数据点划分到某一个超聚类中的信息增益，根据这些权重计算 score，最后取得分最高的超聚类，作为该数据点的归类结果。

相关知识：

1. 香农熵与信息增益：熵定义为信息的期望值，如果待如果待分类的事务可能划分在多个分类之中，则符号x_i的信息定义为

$$\mathrm{l}(x_i) = -log_2 p(x_i)$$

为了计算熵，我们需要计算所有类别所有可能值包含的信息期望值，根据的公式是：

$$H = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

而熵可以用来表征信息的不确定性，这一点可以根据公式进行解释，加入待分类的事物是确切可以肯定被分类到某一类的，根据公式 H=0，表示信息的不确定性为 0，相反，假如事物不能很确切地分类到某一类，H 值就会越大，因此 H 越大表示分类结果的越不能确定。而在算法中，我们需要根据某个数据点被分类到某一个超聚类的得分判断数据点的最终归属，那么就可以使用香农熵进行评估。

2. Jaccard similarity：计算公式为两个集合的交集的大小除以并集的大小，两个越相似的集合的交集越大，得到的相似度就越大，这个很好理解。
3. Ncut 算法：Ncut 算法作为最小割算法的优化，先介绍最小割算法：从图中切去一些边，使图形的源点到终点不连通，当切去的这些边的 cost 最小，即为最小割。然而，最小割算法考虑的是切割图形后子图间的花费最小，而忽略了同一子图内部的耦合度，没有达到整体最优的目的。因而 Ncut 算法在考虑子图间的耦合度的同时也把子图内部的耦合度加入到评测标准，这是因为对于聚类算法而言，除了要保证不同簇之间存在较大差异性以外，也要保证同一簇内部的差异性较小，因此 Ncut 虽然不能保证不同子图间的耦合度最小，但是能达到不同子图间的耦合度较小的同时相同子图内部的耦合度较大，达到聚类的目的。

一个疑问：按照最小割算法和 Ncut 算法的原理，对图形的划分一次只能划分为两个子图，要接着划分的话就要对子图进行划分，那么应该选择哪些子图来划分，同时划分到什么程度算法才停止呢?