Co-clustering on large data arrays

- 目标: 大规模数据数组的 Co-clustering
- 策略: 将数据数组 (2D 的矩阵) 分成小块, 然后合并 sub-co-clusters
- 问题: 划分后, co-cluster 可能会消失 (不包含在任何 sub-co-clusters 中)
- 可能的解决方案: 以不同的方式进行划分并重复该过程; 从统计学角度分析性能
- 假设:
  - A 是大小为 $M \times N$ 的输入矩阵
  - A 划分为 $Q_m \times Q_n$ 块, 每个块的大小为 $B_m \times B_n$, 即 $M = Q_m B_m$, $N = Q_n B_n$
  - sub-co-cluster 的阈值: 最小行数 T_m, 最小列数 T_n
  - 在块 i 中,sub-co-cluster 大小为 $M_i \times N_i$
  - 我们执行 T_p 次划分和重复检测过程
- 分析:
  - 现在考虑一个单个无噪声的 co-cluster, 大小为 $M_s \times N_s$: 如果我们至少有一个 sub-co-cluster, 其大小至少为 $T_m \times T_n$, 那么我们可以通过递归检查行和列中的点来扩大它. 这一步非常简单. 然后我们可以获得大的 co-cluster.
  - 问题是, 如果我们在任何块中都没有检测到任何 sub-co-cluster 会发生什么. 也就是说, 所有 sub-co-cluster 的大小都小于 $T_m \times T_n$.
  - 现在是时候开发有用的 (和相关的) 数学和统计公式了. 我们需要找到在任何块中都没有检测到任何 sub-co-cluster 的概率:

$$P(M_i < T_m, N_i < T_n; \forall i \in \pi_j | Q_m, Q_n, M_s, N_s)$$

  - 我相信这可以通过组合获得. 我们可以以不同的方式 (例如, 一个划分是 $\{\{1,2,3\},\{4,5,6\}\}$, 另一个是 $\{\{1,3,5\},\{2,4,6\}\}$) 对矩阵进行划分, 这可以通过随机采样完成. 现在假设划分是独立的, 在所有划分中都没有检测到 sub-co-cluster 的概率是:

$$\prod_{j=1}^{T_p} P(M_i < T_m, N_i < T_n; \forall i \in \pi_j | Q_m, Q_n, M_s, N_s)$$

  - 检测到 co-cluster 的概率是:

$$1 - \prod_{j=1}^{T_p} P(M_i < T_m, N_i < T_n; \forall i \in \pi_j | Q_m, Q_n, M_s, N_s)$$

    有了这个作为指导, 我们会有想法来设置 T_m, T_n, Q_m, Q_n 和 T_p.

- 任务:
  - 寻找概率
  - 检查理论对单个无噪声 co-cluster 是否正确, co-cluster 有很多不同的大小
  - 嵌入更多无噪声 co-cluster
  - 尝试嘈杂 co-cluster
  - 在模拟和真实例子中试试

我认为如果您能检测到大量椭圆或者找到另一个应用, 就可以写一篇好论文.

如我所说, 要发表一篇好论文, 您需要一个好的想法、好的描述和令人印象深刻的结果. 我相信我们已经有了第一个, 但我们还需要后面的 2 和 3.