基于 GSDPMM 模型改进的聚类消歧方法

1.方法

$$\theta|\alpha \sim GEM(1,\alpha)$$

$$z_d|\theta \sim Mult(\theta) \qquad d = 1,\ldots,D$$

$$\phi_k|\beta \sim Dir(\beta) \qquad k = 1,\ldots,\infty$$

$$d|z_d,\{\phi_k\}_{k=1}^{\infty} \sim p(d|\phi_{z_d})$$

基于折叠吉布斯采样的狄利克雷过程多项式混合模型的主要想法如上图所示,狄利克雷过程多项式混合通过假设有无限个隐形簇，可以完全通过数据学习逼近真实簇的个数。使用折叠吉布斯采样加快了聚类的速度。

算法如下:

Data: 文档向量$\vec{d}$。

Result: 簇的数目K，文档所在簇的向量$\vec{z}$。

begin

//初始化

$K = K_0$ (默认$K_0 = 1$)

将每个簇z的统计量m_z, n_z, n_z^w初始化为零

for $d \in [1, D]$ **do**

//为文档d等概率选择一个簇

$z_d \leftarrow z \sim Mult(1/K)$

$m_z \leftarrow m_z + 1, n_z \leftarrow n_z + N_d$

for $w \in d$ **do**

$n_z^w \leftarrow n_z^w + N_d^w$

//启发式搜索

for $i \in [1, I]$ **do**

for $d \in [1, D]$ **do**

记录文档d当前所在的簇：$z \leftarrow z_d$

$m_z \leftarrow m_z - 1, n_z \leftarrow n_z - N_d$

for $w \in d$ **do**

$n_z^w \leftarrow n_z^w - N_d^w$

if $m_z == 0$ **then**

$K \leftarrow K - 1$

重新排列剩下的非空簇，使其编号为$1, ..., K$

计算文档d选择每个已有簇的概率，以及文档d选择新簇的概率

为文档d按照上述概率随机选择一个簇z

if $z == K + 1$ **then**

$K \leftarrow K + 1$

将簇z的统计量$m_z$fi$n_z$fin_z^w初始化为零

$z_d \leftarrow z, m_z \leftarrow m_z + 1, n_z \leftarrow n_z + N_d$

for $w \in d$ **do**

$n_z^w \leftarrow n_z^w + N_d^w$

2.评价指标

使用了衡量同质性和完整性的 ARI 指标如下

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]}$$

3.超参数设置:

狄利克雷过程中涉及两个参数 gamma 和 beta，分别设为 0.1 和 0.05

4.数据处理:

通过人工标注了一部分数据，作为初始簇的分类标准

5.参考

[1]Jianhua Yin and JianyongWang. 2016. A model-based approach for text clustering with outlier detection. In ICDE. IEEE, 625–636.