



文本处理实践
第五讲

文本聚类

1

文本聚类

- 简介
- 评价指标
- 实验数据集和实验方法

文本聚类 简介

- 聚类的概念

聚类（clustering）：聚类试图将数据集中的样本划分为若干个通常是不相交的子集，每个子集称为一个“簇”（cluster）。

通过这样的划分，每个簇可能对应于一些潜在的类别。这些概念对聚类算法而言事先是未知的，聚类过程仅能自动形成簇结构，簇所对应的含义需要由使用者来把握和命名。

聚类常用于寻找数据内在的分布结构，也可作为分类等其他学习任务的前驱过程。

例如，在一些商业应用中需要对用户类型进行判别，但事先没有定义好的“用户类型”，可以先对用户数据进行聚类，根据聚类结果将每个簇定义为一个类，然后基于这些类训练分类模型，用于判别新用户的类型。

文本聚类 性能度量

- 什么样的聚类结果比较好？
- 同一簇的样本尽可能彼此相似，不同簇的样本尽可能不同。

簇内相似度 (intra-cluster similarity) 高，簇间相似度 (intra-cluster similarity) 低

两类性能度量方法：一类是将聚类结果与某个“参考模型”进行比较，称为“外部指标”。另一类时直接考察聚类结果而不利用任何参考模型，称为“内部指标”。

文本聚类 性能度量

- 外部指标

对数据集

$$D = \{x_1, x_2, \dots, x_m\},$$

通过聚类给出的簇划分为 $C = \{C_1, C_2, \dots, C_k\}$, 参考模型给出的簇划分为 $C^* = \{C^*_1, C^*_2, \dots, C^*_s\}$

相应地, 令 λ 与 λ^* 分别表示与 C 和 C^* 对应的簇标记向量。我们将样本两两配对考虑,

$a = |SS|$, 集合 SS 包含了在 C 中隶属于相同簇且在 C^* 中也隶属于相同簇的样本对。

$$SS = \{(x_i, x_j) \mid \lambda_i = \lambda_j, \lambda_i^* = \lambda_j^*, i < j\}$$

$b = |SD|$, 集合 SD 包含了在 C 中隶属于相同簇但在 C^* 中隶属于不同簇的样本对。

$c = |DS|$

$d = |DD|$

其中, 每个样本对 (x_i, x_j) ($i < j$) 仅能出现在一个集合中, 因此有 $a + b + c + d = m(m-1)/2$

文本聚类 性能度量

- 外部指标
- Jaccard系数 $JC = \frac{a}{a+b+c}$
- FM指数 (简称FMI) $FMI = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$
- Rand指数(简称RI) $RI = \frac{2(a+d)}{m(m-1)}$
- 上述性能度量的结果值在[0,1]区间, 值越大越好。

文本聚类 性能度量

- 内部指标（略）

对数据集 $D = \{x_1, x_2, \dots, x_m\}$,

通过聚类给出的簇划分为 $C = \{C_1, C_2, \dots, C_k\}$,

$\text{dist}(\cdot, \cdot)$ 用于计算两个样本之间的距离。

DBI , 值越小越好

DI , 值越大越好

本周作业

完成文本聚类作业。

对测试数据集得到的聚类结果，完成外部评价指标的计算，并完成实践报告。