

4.10 聚类算法评估

CSDN学院 2017年11月



▶大纲



- 常用聚类算法
- 聚类性能评估
- 案例分析



▶ 聚类质量的评价



- 聚类性能评价方法通常分为外部评价法 (external criterion) 和内部评价法 (internal criterion).
- 外部评价法分析聚类结果与另一参考结果 (reference, 如真实类别) 有多相近
 - Adjust Rand Index (ARI), 互信息 (AMI), V-measure、Fowlkes-Mallows scores
- 内部评价法来分析聚类的本质特点
 - Silhouette Coefficient, Calinski-Harabaz Index



Adjusted Rand Index (ARI)



- 假设有N个样本点,参考类别结果为 $C^* = \{c_1^*, ..., c_L^*\}$
- C为一个聚类结果, $C = \{c_1, ..., c_K\}$
- 计算两个聚类结果C和参考类别结果C*的样本点对的数目

$$N_{11} = \#\{(\mathbf{x}_i \ , \mathbf{x}_j) | \mathbf{x}_i \ , \mathbf{x}_j \in C_k; \ \mathbf{x}_i \ , \mathbf{x}_j \in C_l^*\} \ c$$
与 C *中都是同类别的样本对数 $N_{00} = \#\{(\mathbf{x}_i \ , \mathbf{x}_j) | \mathbf{x}_i \in C_{k_1} \ , \mathbf{x}_j \in C_{k_2}; \ \mathbf{x}_i \in C_{l_1}^* \ , \mathbf{x}_j \in C_{l_2}^*\} \ c$ 与 C *中都是不同类别的样本对数

• Rand指数度量两个标签分配C和C*的吻合程度,定义为: $R = \frac{N_{11} + N_{00}}{\binom{N}{2}}$ 分母:集合中可以组成样本对的对数



Adjusted Rand Index (ARI)



 为了实现"在聚类结果随机产生的情况下,指标应该接近零", Hubert and Arabie (1985)对 Rand index 进行了调整 (减去随机类别结果的RI的期望E(RI)),提出调整兰德系数(Adjusted Rand Index, ARI):

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

- 取值范围为[-1,1]
 - 对两种独立的聚类值为 0, 两种完全相同的聚类值为 1
- 值越大意味着聚类结果与真实情况越吻合

▶基于互信息的分数(AMI)

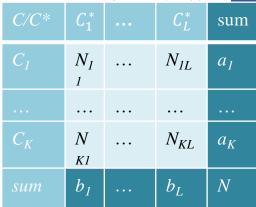
- AMI中用到的记号和ARI一样,只是将Rand指标换成互信息
- 回忆信息中关于熵 / 信息的定义:
 - $\Rightarrow N_{kl}$ 表示同时属于C中 C_k 和 C^* 中 C_l^* 的样本数目, $a_k = \sum_{l=1}^L N_{kl}$ 为C中 C_k 中的样本数目, $b_l = \sum_{k=1}^K N_{kl} 为 C^* + C_l^*$ 中的样本数目,则 $H(C) = \sum_{k=1}^{K} p(k) log p(k)$, where $p(k) = \frac{a_k}{N}$

$$H(C) = \sum_{k=1}^{l} p(k) log p(k)$$
, where $p(k) = \frac{\kappa}{N}$
 $H(C^*) = \sum_{l=1}^{L} p'(l) log p'(l)$, where $p'(l) = \frac{b_l}{N}$

C和参考类别结果C*之间的<u>互信息</u>定义为: $MI(C,C^*) = \sum_{k=1}^{K} \sum_{l=1}^{L} p(k,l) \log \frac{p(k,l)}{p(k)p'(l)}, \text{ where } p(k,l) = \frac{N_{kl}}{N}$

▶ 调整的互信息 (Adjusted Mutual Informations)

• AMI对互信息的调整与ARI对RI的调整类似 $AMI = \frac{MI - E(MI)}{\max(H(C), H(C^*)) - E(MI)}$



• MI的期望为
$$E(MI) = \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{N_{ij}=(a_k+b_l-N)^+}^{min(a_k,b_l)} \frac{N_{ij}}{N} \log \frac{N \times N_{kl}}{a_k b_l} \times \frac{a_k! \, b_l! \, (N-a_k)! \, (N-b_l)!}{N! \, N_{kl}! \, (a_k-N_{kl})! \, (b_l-N_{kl}) (N-a_k-b_l+N_{kl})!}$$

- AMI的取值范围为[0,1]
 - 对两种独立的聚类值为0,两种完全相同的聚类值为1



and V-measure

Homogeneity, completeness



$$N_{K1}$$

 N_{11}

$$egin{array}{c} N_{KL} \ b_L \end{array}$$

 N_{IL}

. . .

 a_K

sum

• 同质性 (homogeneity)
$$h = 1 - \frac{H(C|C^*)}{H(C)}$$

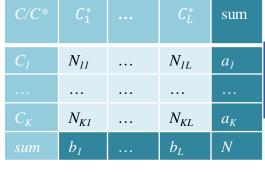
- 其中
$$H(C)$$
为聚类结果 C 的熵:
$$H(C) = \sum_{k=1}^{K} p(C_k) \log p(C_k) = \sum_{k=1}^{K} \frac{N_k}{N} \log \frac{N_k}{N}$$

$$H(C/C^*)$$
为给定簇分配 C^* 条件下的类 C 的条件熵:



 $H(C \mid C^*) = \sum_{k=1}^{K} \sum_{l=1}^{L} p(C_k, C_l^*) \log \frac{p(C_l^*)}{p(C_k, C_l^*)} = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{N_{kl}}{N} \log \frac{N_{kl}}{N_l}$

Homogeneity, completeness and V-measure



• 完整性 (completeness):同类别样本被归类到相同簇中 $H(C^*|C)$

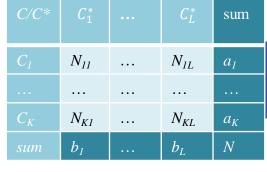
$$c = 1 - \frac{H(C^* \mid C)}{H(C^*)}$$

• V-measure为均一性和完整性的调和平均:

$$v = 2\frac{h \times c}{h + c}$$



Fowlkes-Mallows scores



- Fowlkes-Mallows score (FMI) 为成对精度 (precision) 和召回率 (recall) 的几何均值
 - -精度: $p = N_{kl}/a_k$
 - 召回率: $r = N_{kl}/b_l$

$$- FMI = \sqrt{p \times r} = \sqrt{\frac{N_{kl} \times N_{kl}}{a_k \times b_l}} = \frac{N_{kl}}{\sqrt{a_k \times b_l}}$$



▶聚类质量的评价



- 聚类性能评价方法通常分为外部评价法 (external criterion) 和内部评价法 (internal criterion).
- 外部评价法分析聚类结果与另一参考结果 (reference, 如真实类别) 有多相近
 - Adjust Rand Index (ARI), 互信息 (AMI), V-measure、Fowlkes-Mallows scores
- 内部评价法来分析聚类的本质特点,又分为绝对评价和相对评价法,常用的方法有
 - Silhouette Coefficient, Calinski-Harabaz Index



▶轮廓系数 (Silhouette coefficient)



类内散度

类间散度

- 轮廓系数(侧影法)适用于实际类别信息未知的情况
- 对于其中的一个样本点 *i* , 记:
 - a(i): 样本点i到与其所属簇中其它点的平均距离
 - $-\bar{d}(i,C)$: 样本点i到其他类 $C(C \neq C_i)$ 内所有点的平均距离
 - -b(i): 所有 $\bar{d}(i,C)$ 的最小值
- 则样本点i 的轮廓系数为: $s(i) = \frac{b(i) a(i)}{\max\{a(i), b(i)\}}$



平均Silhouette值为: $\overline{s} = \frac{1}{N} \sum_{i=1}^{N} s(i)$

▶轮廓系数 (Silhouette coefficient)



- $-1 \le s(i) \le 1$ (当 $a(i) \ll b(i)$, s(i)接近1)
 - 由a(i)的定义可知:小的a(i)意味着点i匹配该类非常好,而大的b(i)意味着点i匹配其他类很差,从而s(i)靠近1表明点i的聚类合适
- s(i)靠近-1表明点i被聚类到相邻类中更合适
- s(i)靠近0表明点i在两个类的交集处

- 可使用轮廓系数估计聚类中的类的数目
 - $-\overline{s} > 0.5$ 表明聚类合适
 - $-\overline{s} < 0.2$ 表明数据不存在聚类特征



► CH索引 (Calinski-Harabaz Index)

- N_{II} ... N_{IL} a_1 N_{KL} a_K ... a_K ... a_K
- CH索引也适用于实际类别信息未知的情况
- 下面以K-means算法为例,可推广到其他算法。
- 给定聚类数目K, 类内散度为: $W(K) = \sum_{k=1}^{K} \sum_{C(j)=k} \|\mathbf{x}_j \overline{\mathbf{x}}_k\|^2$
- 类间散度: $B(K) = \sum_{k=1}^{K} a_k \|\overline{\mathbf{x}}_k \overline{\mathbf{x}}\|^2$
- 则CH索引为 $CH(K) = \frac{B(K)(N-K)}{W(K)(K-1)}$

计算快

sum









THANK YOU



