

# 中文信息熵分析

ZY2207304-范想明

信息熵是英国数学家克劳德·香农于 1948 年提出的一个表征符号系统中单位符号平均信息量的指标，并给出了一个十分简洁的公式：

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

式中： $P(x_i)$ ——某符号系统中符号  $x_i$  出现的概率

假设一个句子中由特定的字或词 ( $x_i, i \in 1, 2, \dots, n$ ) 组成，则其概率分布为：

$$P(x_1, x_2, \dots, x_n) = P(x_1)P(x_2 | x_1) \dots P(x_n | x_1, x_2, \dots, x_{n-1}) \quad (2)$$

以分析单个符号（字或词）为例，一般其分布构建简化为  $n$  元语言模型，即将每个单元的概率只与前 ( $n-1$ ) 个单元有关，其中一元模型可表示为：

$$P(x_1, x_2, \dots, x_n) = P(x_1)P(x_2) \dots P(x_n) \quad (3)$$

以此类推，二元、三元模型可分别表示为：

$$P(x_1, x_2, \dots, x_n) = P(x_1)P(x_2 | x_1) \dots P(x_n | x_{n-1}) \quad (4)$$

$$P(x_1, x_2, \dots, x_n) = P(x_1, x_2)P(x_3 | x_2, x_1) \dots P(x_n | x_{n-2}, x_{n-1}) \quad (5)$$

因此，根据上述简化模型，一元、二元和三元模型下，其信息熵可分别表示为：

$$H(X) = -\sum_{i=1}^n P(x_i) \log_2 P(x_i) \quad (6)$$

$$H(X | Y) = -\sum_{i=1}^n P(x_i | y_i) \log_2 P(x_i | y_i) \quad (7)$$

$$H(X | Y, Z) = -\sum_{i=1}^n P(x_i, y_i, z_i) \log_2 P(x_i | y_i, z_i) \quad (8)$$

其中联合概率和条件概率的计算根据《数理统计》中的公式进行计算即可。