第二章离散信源信息度量

第五节信源的冗余度

陈兴同

中国矿业大学 数学学院

2021 年 8 月

内容提要

熵率与字符依赖性

信源的熵率实质上表达了一个信源字符所包含或能携带的平均信息量（称为实在信息）。如果考虑到信源发出的长消息中各字符之间的相互依赖性，则熵率随着消息序列的长度增加而减小即

$$H_\infty(X) \leq \cdots \leq H(X_n|X_1X_2\cdots X_{n-1}) \leq \cdots \leq H(X_2|X_1) \leq H(X_1).$$

字符之间的这种依赖关系通常是由于语法要求、传统习惯等约束而形成的，约束越大，依赖性就越强，每个字符能携带的信息量就越少。每个字符所能携带的最大信息量即 $\log N$ 与实在信息量即熵率 $H_\infty(X)$ 之间的差称为信源的冗余。

冗余定义2.5.1

对有 N 个字符的信源 X，它的熵率为 $H_\infty(X)$，称

$$R = \log N - H_\infty(X),$$

***为信源的冗余**；而称*

$$\eta = 1 - \frac{H_\infty(X)}{\log N} = \frac{R}{\log N}.$$

***为信源的冗余度**。*

信源的这种冗余完全可以通过压缩编码去掉，然后再进行存储或传输。

例题 2.5.1：

以英文字符构成的信源说明信源的冗余度。

解：书面英语中包含大量的冗余。考虑一个只由 26 个英文字母和空格作为字符空间构成的信源。仙农将这种信源按字母和单词模型分别估计了信源的熵率 [20].

（1）如果把信源看成是离散无记忆的，即各个字符是相互独立互不依赖的，则这个信源的最大熵为 $H_0 = 4.7$bits，但实际上各个字母出现的概率不同（可以查到这种概率分布表），这时信源的熵为 $H_1 = 4.03$bits。

（2）如果考虑两个字母间的相互依赖关系把信源看成为马氏信源，则熵率为 $H_2 = 3.6$bits。

续解：

（3）如果考虑三个字符间的依赖关系，则信源可看成二阶马氏信源，熵率为 $H_3 = 3.3$bits。

（4）如果考虑四个字符间的依赖关系，把信源看成为三阶马氏信源，则熵率为 $H_4 = 2.8$bits。

（5）如果考虑能构成有意义句子的字母，则熵率会更小。申农估计在这种情况下熵率 $H_\infty = 1.4$bits！可见口头或书面英文中包含了大量的冗余！