

第 8 讲 赫夫曼码与最优码

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

1952 年赫夫曼提出一种构造最佳码的方法称为赫夫曼码。

下面首先给出二元赫夫曼码的编码方法，它的编码步骤如下：

1. 将 q 个信源符号按概率分布 $P(s_i)$ 的大小，以递减次序排列起来，设

$$p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_q.$$

2. 将 0 和 1 分别分配给概率最小的两个信源符号，并将这两个概率最小的信源符号合并成一个新符号，并用这两个最小概率之和作为新符号的概率，从而得到只包含 $q - 1$ 个符号的新信源，称为信源 S 的缩减信源 S_1 .
3. 把缩减信源 S_1 的符号仍按概率大小以递减依次序排列，再将其最后两个概率最小的符号合并成一个新符号，并分别用 0 和 1 码符号表示，这样得到了 $q - 2$ 个符号的缩减信源 S_2 .
4. 依次继续下去，直至缩减信源最后只剩两个符号为止。将这最后两个符号分别用 0 和 1 码符号表示。最后这两个符号的概率之和必为 1。然后从最后一级缩减信源开始，依编码路径由后向前返回，就得到各信源符号所对应的码符号序列，即得对应的码字。

二元赫夫曼码

 r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

上面我们讨论了二元赫夫曼码，它的编码方式同样可以推广到 r 元编码中来。不同的只是把 r 个符号（概率最小的）合并成一个新信源信号，并分别用 $0, 1, \dots, (r-1)$ 等码元表示。为了使短码得到充分利用，使平均码长最短，必须使最后一步的缩减信源有 r 个信源符号。因此对于 r 元编码，信源 S 的符号个数 q 必须满足

$$q = (r-1)\theta + r.$$

式中， θ 表示缩减的次数， $(r-1)$ 为每次缩减所减少的信源符号个数。对于二元码，信源符号个数 q 必须满足

$$q = \theta + 2.$$

可见对于二元码， q 等于任意正整数时，总能找到一个 θ 满足上式。而对于 r 元码，则不一定。此时我们不妨虚设 t 个信源符号 $s_{q+1}, s_{q+2}, \dots, s_{q+t}$ 并使它们对应的概率为零，即 $p_{q+1} = p_{q+2} = \dots = p_{q+t} = 0$ 。此时， $q+t = (r-1)\theta + r$ 。

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-
Elias 编码

香农码的竞争最优性

赫夫曼码的优点:

- ▶ 赫夫曼码的编码方法保证了概率大的符号对应于短码，概率小的符号对应于长码，充分利用了短码；
- ▶ 缩减信源的最后两个码字总是最后一位不同，从而保证了赫夫曼码是即时码。

赫夫曼码的缺点:

- ▶ 每个信源符号所对应的码长不同. 一般情况下, 信源符号以恒定速度输出, 信道也是恒速传输的. 通过编码后, 会造成编码输出每秒的比特数不是常量, 因而不能直接由信道来传送.
- ▶ 信源符号与码字之间不能用某种有规律的数学方法对应起来.

从赫夫曼码的编码过程可以看出，赫夫曼码的编码方式得到的码一定是即时码，因为这种编码方法不会使任一码字的前缀为码字。

注意到赫夫曼码编码方法得到的码并非使唯一的。首先因为，每次对缩减信源最后两个概率最小的符号，用 0 和 1 码是任意的，所以可以得到不同的码。但它们只是码字的具体形式不同，而其码长 l_i 不变，平均码长 $\bar{L}$ 也不变，所以没有本质差别。其次，若当缩减信源中缩减合并的符号的概率与其他信源符号概率相同时，从编码方法上来说，它们概率次序的排列哪个在前哪个在后是没有区别的，但得到的码使不同的赫夫曼码。对于这两种码，他们的码长 l_i 各不同，然而平均码长 $\bar{L}$ 是相同的。

那么，在这些不同的码中，选哪一个码好呢？我们引进码字长度 l_i 偏离平均长度 $\bar{L}$ 的方差 σ^2 ，即

$$\sigma^2 = E[(l_i - \bar{L})^2] = \sum_{i=1}^q P_i (l_i - \bar{L})^2.$$

当方差 σ^2 比较小时，码序列长度变化比较小，相对来说效果会更好。

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

引理 4.1

对于任意一个分布, 必然存在满足如下性质的一个即时最优码 (即有最小期望长度):

1. 其长度序列与按概率分布列排列的次序相反, 即, 若 $p_j > p_k$, 则 $l_j \leq l_k$.
2. 最长的两个码字具有相同的长度.
3. 最长的两个码字在最后一位上有所差别, 且对应于两个最小可能发生的字符.

证明.

考虑一个最优码 C_m :

- 若 $p_j > p_k$, 则 $l_j \leq l_k$. 此时通过交换码字即可得此结论.
设 C'_m 为将 C_m 中的码字 j 和 k 交换所得到的编码, 则

$$\begin{aligned} L(C'_m) - L(C_m) &= \sum p_i l'_i - \sum p_i l_i \\ &= p_j l_k + p_k l_j - p_j l_j - p_k l_k \\ &= (p_j - p_k)(l_k - l_j). \end{aligned}$$

但 $p_j - p_k > 0$, 由于 C_m 是最优的, 可得 $L(C'_m) - L(C_m) \geq 0$, 故必有 $l_k \geq l_j$. 从而最优码本身 C_m 必定满足性质 1.

证明.

- ▶ 最长的两个码字具有相同的长度.
通过修建码字获得结论. 如果两个码字最长码长度不相同, 那么将较长的码字最后一位删除, 它仍可保持前缀性质, 但此时具有更短的期望码字长. 因此, 最长的两个码字长度必定相等. 由性质 1 我们知, 最长的码字对应于那些最小可能发生的符号信源字符.
- ▶ 两个最长码仅在最后一位有所差别, 并且分别对应两个最小可能发生的信源字符.
并非所有的最优码都满足这个性质, 但通过重排可以获得满足该性质的最优码. 如果存在长度最长的码字, 则删除码字的最后一位, 所得的码字仍然满足前缀性质. 从而期望的码字长度有所减小, 这与编码的最优性矛盾. 因此, 在任何一个最优编码中, 最大长度码字有兄弟.

总之, 我们已证明: 若 $p_1 \geq p_2 \geq \cdots \geq p_m$, 则存在长度列为 $l_1 \leq l_2 \leq \cdots \leq l_{m-1} = l_m$ 的一个最优码, 且码字 $C(x_{m-1})$ 和 $C(x_m)$ 仅最后一位有所区别. □

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

满足上述引理的最优码我们称之为典则码. 对于 m 个字符上的概率密度函数 $\mathbf{p} = (p_1, p_2, \dots, p_m)$, $p_1 \geq p_2 \geq \dots \geq p_m$, 我们定义其 $m-1$ 个字符上的赫夫曼合并为 $\mathbf{p}' = (p_1, p_2, \dots, p_{m-1} + p_m)$. 用 $C_{m-1}^*(\mathbf{p}')$ 表示 $\mathbf{p}'$ 的最优码, 而用 $C_m^*(\mathbf{p})$ 表示 $\mathbf{p}$ 的典则最优码.

定理 4.2

赫夫曼码是最优的, 即如果 C^* 为赫夫曼码而 C' 为其他码, 则有 $L(C^*) \leq L(C')$.

证明.

我们利用归纳法, 对信源的字符个数做归纳. 当信源有一个和两个字符时, 显然赫夫曼码是最优的. 下面我们假设对于 $m-1$ 个字符的信源, 赫夫曼码是最优的. 对于 m 个字符上的概率密度函数

$\mathbf{p} = (p_1, p_2, \dots, p_m), p_1 \geq p_2 \geq \dots \geq p_m$, 我们定义其 $m-1$ 个字符上的赫夫曼合并为 $\mathbf{p}' = (p_1, p_2, \dots, p_{m-1} + p_m)$. 用 $C_{m-1}^*(\mathbf{p}')$ 表示 $\mathbf{p}'$ 的最优码, 而用 $C_m^*(\mathbf{p})$ 表示 $\mathbf{p}$ 的典则最优码. 记 $L^*(\mathbf{p})$ 和 $L^*(\mathbf{p}')$ 为对应的概率密度函数所对应的最优码平均码长.

基于 $\mathbf{p}'$ 的赫夫曼码我们可以构造一个对应 m 字符信源的编码: 取 C_{m-1}^* 中的权重为 $p_{m-1} + p_m$ 的码字, 对其进行扩展, 在尾部加 0 形成字符 $m-1$ 相应码字, 加 1 形成字符 m 的相应码字. 由此得到的新码是对应的赫夫曼码, 码字平均长度为

$$L(\mathbf{P}) = L^*(\mathbf{p}') + p_{m-1} + p_m.$$

证明.

另外, 我们从 $\mathbf{p}$ 的典则码出发, 将两个最小概率 p_{m-1} 和 p_m 对应的字符 $m-1$ 和 m 合并, 可以构造关于分布 $\mathbf{p}'$ 的新码, 其平均长度为

$$L(\mathbf{p}') = L^*(\mathbf{p}) - p_{m-1} - p_m.$$

于是

$$(L(\mathbf{p}') - L^*(\mathbf{p}')) + (L(\mathbf{p})) - L^*(\mathbf{p}) = 0$$

而由最优码有最短的平均码字, 故我们有 $L(\mathbf{p}) = L^*(\mathbf{p})$. 从而我们知上面得到的 m 元字符信源的赫夫曼码是最优码. $\square$

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

不失一般性, 假定取 $\mathcal{X} = \{1, 2, \dots, m\}$. 假设对所有的 x , 有 $p(x) > 0$. 定义累积分布函数 $F(x)$ 为

$$F(x) = \sum_{a \leq x} p(a).$$

考虑修正的累积分布函数

$$\bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2}p(x).$$

由于所有的概率值是正的, 若 $a \neq b$, 则 $\bar{F}(a) \neq \bar{F}(b)$. 若已知 $\bar{F}(x)$, 则可以确定 x . 故 $\bar{F}(x)$ 可以作为 x 的编码. 但是在一般情况下, $\bar{F}(x)$ 是一个无理数, 其需要无限多比特才能表示. 所以使用 $\bar{F}(x)$ 的精确值作为对 x 的编码并非切实可行. 那么如果我们要使用近似值, 那么需要精确到什么程度呢? 假定将 $\bar{F}(x)$ 舍入取 $l(x)$ 位 (记为 $\lfloor \bar{F}(x) \rfloor_{l(x)}$). 于是, 取 $\bar{F}(x)$ 的前 $l(x)$ 位作为 x 的码. 由舍入定义, 可得

$$\bar{F}(x) - \lfloor \bar{F}(x) \rfloor_{l(x)} < \frac{1}{2^{l(x)}}.$$

若 $l(x) = \lceil \log \frac{1}{p(x)} \rceil + 1$, 则

$$\frac{1}{2^{l(x)}} < \frac{p(x)}{2} = \bar{F}(x) - F(x-1).$$

于是 $\lfloor \bar{F}(x) \rfloor_{l(x)}$ 和 $\lfloor \bar{F}(x-1) \rfloor_{l(x)}$ 不同. 从而, 使用 $l(x)$ 比特足以表示 x . 这里除了要求码字和字符一一对应之外, 还需要码字集是无前缀的. 为验证该编码是否为前缀码是否为前缀码, 考虑每个码字 $z_1 z_2 \cdots z_l$, 注意到它实际上代表的不是一个点, 而是一个区间 $[0.z_1 z_2 \cdots z_l, 0.z_1 z_2 \cdots z_l + \frac{1}{2^l}]$. 码是无前缀的当且仅当码字对应的区间互不相交.

下面我们来证明上述码字集合无前缀. 对应任意码字的区间长度为 $2^{-l(x)}$, 由 (20) 可知所有区间长度均小于 x 对应的阶梯高度的 $1/2$. 区间的下端位于对应阶梯的下一半中, 于是区间的上端位于对应阶梯的顶部之下, 故而在累积分布函数之中, 任一码字对应的区间都真包含于相应字符所对应的阶梯中. 所以不同码字对应的区间不相交, 于是我们知此码是无前缀的. 注意, 该程序没有要求字符按其概率大小顺序排列.

我们可以计算编码的期望长度:

$$L = \sum_x p(x)l(x) = \sum_x p(x)(\lceil \log \frac{1}{p(x)} \rceil + 1) < H(X) + 2.$$

因此, 该编码方案的期望码长不会超过熵值 2 比特.

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

定理 6.1

设 $l(x)$ 为香农码的相应码字长度, 而 $l'(x)$ 表示其他唯一可译码的相应码字长度. 则

$$P(l(X) \geq l'(X) + c) \leq \frac{1}{2^{c-1}}.$$

$$\begin{aligned}
 P(l(X) \geq l'(X) + c) &= P(\lceil \log \frac{1}{p(X)} \rceil \geq l'(X) + c) \\
 &\leq P(\log \frac{1}{p(X)} \geq l'(X) + c - 1) \\
 &= P(p(X) \leq 2^{-l'(X)-c+1}) \\
 &= \sum_{x: p(x) \leq 2^{-l'(x)-c+1}} p(x) \\
 &\leq \sum_{x: p(x) \leq 2^{-l'(x)-c+1}} 2^{-l'(x)-(c-1)} \\
 &\leq \sum_x 2^{-l'(x)} 2^{-(c-1)} \\
 &\leq 2^{-(c-1)}.
 \end{aligned}$$

最后一步我们用到了 Kraft 不等式.

定理 6.2

对二进制概率密度函数 $p(x)$, 设 $l(x) = \log \frac{1}{p(x)}$ 为信源的二维香农码的码字长度, $l'(x)$ 为信源任何其他唯一可译二元码的码字长度. 则

$$P(l(X) < l'(X)) \geq P(l(X) \geq l'(X)).$$

当且仅当对所有的 x , 有 $l'(x) = l(x)$ 等号成立. 于是码长分配 $l(x) = \log \frac{1}{p(x)}$ 是唯一竞争最优的.

证明.

定义函数 $\text{sgn}(t)$ 如下:

$$\text{sgn}(t) = \begin{cases} 1 & \text{当 } t > 0 \\ 0 & \text{当 } t = 0 \\ -1 & \text{当 } t < 0 \end{cases}$$

我们不难看出

$$\text{sgn}(t) \leq 2^t - 1, \quad t = 0, \pm 1, \pm 2, \dots$$

证明.

于是我们可以得到

$$\begin{aligned}
P\{l'(X) < l(X)\} - P\{l'(X) > l(X)\} &= \sum_{x:l'(x) < l(x)} p(x) - \sum_{x:l'(x) > l(x)} p(x) \\
&= \sum_x p(x) \operatorname{sgn}(l(x) - l'(x)) \\
&= E \operatorname{sgn}(l(X) - l'(X)) \\
&\leq \sum_x p(x) (2^{l(x)-l'(x)} - 1) \\
&= \sum_x 2^{-l(x)} (2^{l(x)-l'(x)} - 1) \\
&= \sum_x 2^{-l'(x)} - \sum_x 2^{-l(x)} \\
&= \sum_x 2^{-l'(x)} - 1 \leq 1 - 1 = 0.
\end{aligned}$$

讨论上面不等式等号成立的条件我们知当且仅当对所有的 x , 有 $l'(x) = l(x)$ 等号成立.

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

推论 6.3

对于非二进的概率密度函数,

$$E \operatorname{sgn}(l(x) - l'(x) - 1) \leq 0.$$

其中 $l(x) = \lceil \log \frac{1}{p(x)} \rceil$, $l'(x)$ 为信源其他任何一个编码.

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性

例 7.1

设 $Z_1, Z_2, \dots$ 为独立同分布的随机变量, 服从 $\{0, 1\}$ 上的均匀分布. 对于 $1 \leq i \leq n$, 设

$$X_i = \sum_{j=1}^i Z_j.$$

求 $I(X_1; X_2, X_3, \dots, X_n)$.

证明.

首先注意到 $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ 构成一个马尔可夫链. (思考: 为什么?)

由互信息的链式法则, 我们有

$$\begin{aligned} I(X_1; X_2, X_3, \cdots, X_n) &= \sum_{i=2}^n I(X_1; X_i | X_2, \cdots, X_{i-1}) \\ &= I(X_1; X_2). \\ &= I(Z_1; Z_1 + Z_2) \\ &= H(Z_1 + Z_2) - H(Z_1 + Z_2 | Z_1) \\ &= \frac{3}{2} - 1 = 1/2 \text{ 比特}. \end{aligned}$$

□

例 7.2

设 $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ 构成一个马尔科夫链. 证明:

$$I(X_1; X_3) + I(X_2; X_4) \leq I(X_1; X_4) + I(X_2; X_3).$$

证明.
注意到

$$\begin{aligned} I(X_1; X_3, X_4) &= I(X_1; X_3) + I(X_1; X_4|X_3) \\ &= I(X_1; X_4) + I(X_1; X_3|X_4) \end{aligned}$$

于是

$$I(X_1; X_4) - I(X_1; X_3) = I(X_1; X_4|X_3) - I(X_1; X_3|X_4).$$

类似地,

$$I(X_2; X_3) - I(X_2; X_4) = I(X_2; X_3|X_4) - I(X_2; X_4|X_3).$$

从而我们有

$$\begin{aligned} &I(X_1; X_4) + I(X_2; X_3) - I(X_1; X_3) + I(X_2; X_4) \\ &= I(X_1; X_4|X_3) - I(X_1; X_3|X_4) + I(X_2; X_3|X_4) - I(X_2; X_4|X_3). \end{aligned}$$

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-
Elias 编码

香农码的竞争最优性

证明.
又由

$$\begin{aligned} I(X_1; X_2; X_4|X_3) &= I(X_1; X_4|X_3) + I(X_2; X_4|X_1, X_3) \\ &\quad I(X_2; X_4|X_3) + I(X_1; X_4|X_2, X_3), \end{aligned}$$

以及

$$\begin{aligned} I(X_1, X_2; X_3|X_4) &= I(X_1; X_3|X_4) + I(X_2; X_3|X_1, X_4) \\ &= I(X_2; X_3|X_4) + I(X_1; X_3|X_2, X_4), \end{aligned}$$

我们知

$$\begin{aligned} &I(X_1; X_4) + I(X_2; X_3) - I(X_1; X_3) + I(X_2; X_4) \\ &= I(X_1; X_4|X_2, X_3) - I(X_2; X_4|X_1, X_3) + I(X_2; X_3|X_1, X_4) - I(X_1; X_3|X_2, X_4) \\ &= I(X_2; X_3|X_1, X_4) \geq 0, \end{aligned}$$

从而得证.

二元赫夫曼码

r 元赫夫曼码

赫夫曼码的相关讨论

赫夫曼码的最优性

Shannon-Fano-Elias 编码

香农码的竞争最优性