PSI3531 - Processamento de Sinais Aplicado
23/10/2018 Magno T. M. Silva

Uma introdução à teoria da informação

Thus, we may have knowledge of the past but cannot control it; we may control the future but not have knowledge of it. Claude E. Shannon

Um pouco após a Segunda Guerra Mundial, Claude Shannon (1948) e Norbert Wiener (1949) estabeleceram novos conceitos que formaram a base da moderna teoria de comunicações (CARLSON, 1975). Wiener estava mais interessado em obter a melhor estimativa do sinal transmitido a partir do sinal recebido afetado de ruído. Shannon, por sua vez, estava interessado em representar adequadamente as mensagens produzidas por uma fonte diante das limitações físicas do sistema de comunicação. O estudo de Shannon deu origem à *Teoria da Informação* baseada em três conceitos: medida de informação, capacidade do canal de comunicação para transmitir informação e codificação como forma de utilizar a máxima capacidade do canal. Esses conceitos levaram ao seguinte teorema fundamental da teoria da informação (CARLSON, 1975):

Considerando uma fonte de informação e um canal de comunicação, então existe uma técnica de codificação tal que a informação pode ser trasmitida pelo canal com qualquer taxa menor que a capacidade do mesmo e com uma taxa de erros arbitrariamente pequena devido ao ruído.

Os conceitos de medida de informação, entropia e codificação de Huffman são abordados a seguir.

1 Incerteza, informação e entropia

A saída emitida por uma fonte de informação (dados, voz, vídeo, etc.) pode ser modelada como uma variável aleatória S observada a cada unidade de tempo (intervalo de sinalização) e cujos símbolos pertencem ao alfabeto fixo e finito

$$\mathcal{S} = \{s_0,\ s_1,\ \ldots, s_{K-1}\} \tag{1}$$

com probabilidades

$$P(S = s_k) = p_k, \qquad k = 0, 1, \ldots, K-1. \tag{2}$$

Esse conjunto de probabilidades satisfaz a condição

$$\sum_{k=0}^{K-1} p_k = 1. \tag{3}$$

Assumem-se que os símbolos emitidos pela fonte durante intervalos de sinalização sucessivos são estatisticamente independentes.

Considere o evento $S = s_k$ que descreve a emissão do símbolo s_k pela fonte com probabilidade p_k definida na Equação (2). Se $p_k = 1$ e $p_i = 0$ com $i \neq k$, então não há "incerteza", não há "surpresa" e consequentemente não há "informação" quando o símbolo s_k é emitido porque neste caso se conhece de antemão a mensagem originária da fonte. No entanto, se os símbolos são emitidos com probabilidades diferentes e $p_k \ll p_i$ com $i \neq k$, então há mais surpresa e consequentemente mais informação quando s_k é emitido em comparação com a emissão dos demais símbolos s_i. Desta forma, há uma relação entre as palavras *incerteza*, *surpresa* e *informação*. Antes da ocorrência do evento $S = s_k$ há uma certa incerteza, quando ele ocorre há uma certa surpresa e depois da ocorrência há uma certa informação. A quantidade de informação está relacionada com o inverso da probabilidade de ocorrência.

Define-se a quantidade de informação obtida após a observação do evento $S = s_k$ como

$$I(s_k) = \log\left(\frac{1}{p_k}\right). \tag{4}$$

A base do logaritmo usualmente escolhida nessa equação é a base 2 o que resulta na quantidade de informação expressa em *bits*. Se $p_k = 1/2$ obtém-se $I(s_k) = 1$ bit (*bi*nary digi*t*).

A quantidade de informação $I(s_k)$ produzida pela fonte durante o intervalo de sinalização depende do símbolo s_k emitido em cada instante de tempo. Desta forma, $I(s_k)$ é uma variável aleatória discreta que toma valores $I(s_0),\ I(s_1),\ \ldots,\ I(s_{K-1})$ com probabilidades $p_0,\ p_1,\ \ldots,\ p_{K-1}$, respectivamente. A informação média da fonte, considerando o alfabeto de símbolos $\mathcal{S}$, é dada por

$$H(\mathcal{S}) = \mathrm{E}\{I(s_k)\} = \sum_{k=0}^{K-1} p_k I(s_k) = -\sum_{k=0}^{K-1} p_k \log p_k. \tag{5}$$

A quantidade $H(\mathcal{S})$ é chamada de *entropia* da fonte dado o alfabeto $\mathcal{S}$. Essa é a medida da informação média da fonte por símbolo.

Para o alfabeto binário $\{+1, -1\}$ com probabilidades p e $1-p$, a entropia, denotada por $H_b(p)$, é dada por

$$H_b(p) = -p\log(p) - (1-p)\log(1-p). \tag{6}$$

Um gráfico da entropia para um alfabeto binário é apresentado na Figura 1. Observando essa Figura 1 verifica-se que a entropia é nula quando $p = 0$ e $p = 1$. De fato, considerando uma fonte que possui um alfabeto $\mathcal{S}$ com K símbolos, $H(\mathcal{S}) = 0$ se e somente se $p_k = 1$ e $p_i = 0$ com $i = 0,\ 1,\ \ldots,\ K-1$ e $i \neq k$. Neste caso não há incerteza e portanto não há informação. Voltando para a Figura 1, a entropia é máxima quando $p = 1/2$. Esta é a situação de maior incerteza. Generalizando para um alfabeto de K símbolos, a entropia é máxima e igual a $\log K$ quando todos os símbolos tiverem probabilidades iguais a $1/K$ que corresponde ao caso de maior incerteza. Além dessas, a entropia tem inúmeras propriedades interessantes que mostram que ela é uma medida razoável de informação. Para o leitor interessado recomendam-se as referências (SHANNON, 1948) e (HAYKIN, 2001).

A entropia de uma fonte estabelece um limite fundamental sobre o número de bits necessários para a completa recuperação da mesma. Em outras palavras, para recuperar a informação sem erro, cada saída da fonte deve ter em média um número de bits perto de $H(\mathcal{S})$ mas não menor que $H(\mathcal{S})$. Esse é o princípio básico da codificação da fonte abortada a seguir.

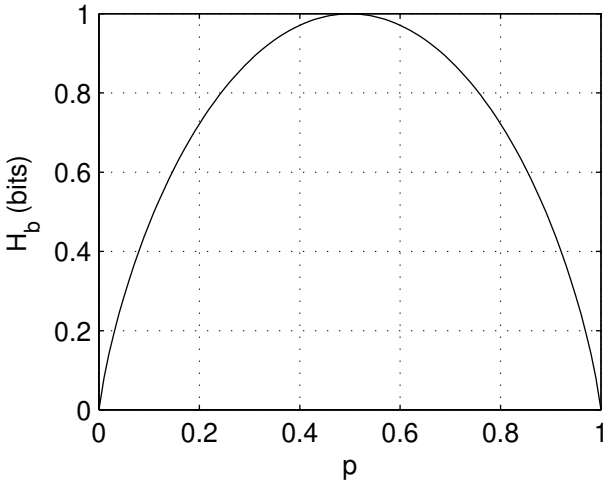

Figura 1: Gráfico da função de entropia binária.

2 Codificação da fonte

Um importante problema em comunicações é a representação eficiente dos dados gerados por uma fonte discreta. O processo que trata dessa representação é chamado de codificação da fonte. Para que tal codificação seja eficiente é necessário conhecer as estatísticas dos símbolos emitidos. Em particular, palavras-código mais longas são atribuídas às saídas da fonte menos prováveis e palavras-código mais curtas às mais prováveis. Neste caso, fala-se de um código de comprimento variável. Um exemplo é o *Código Morse* em que as letras do alfabeto e os números são codificados utilizando pontos "$\cdot$" e traços "$-$". Na língua inglesa, a letra E é mais frequente que a letra Q por exemplo, então o *Código Morse* atribui um simples ponto "$\cdot$" à letra E e "$- \; - \; \cdot \; -$" à letra Q.

Inicialmente deseja-se codificar a fonte de forma eficiente a fim de satisfazer:

1. As palavras-código produzidas pelo codificador são binárias;
2. O código da fonte é unicamente decodificável. Desta forma, a sequência de símbolos originalmente emitida pode ser reconstruída perfeitamente a partir da sequência codificada.

Seja uma fonte com símbolos pertencentes ao alfabeto $\mathcal{S} = \{s_0, \; s_1, \; \ldots, s_{K-1}\}$ com probabilidades $P(S = s_k) = p_k, \quad k = 0, 1, \ldots, K-1$, codificada com palavras-código de comprimento $\{l_0, \; l_1, \; \ldots, l_{K-1}\}$ medido em bits. Define-se o comprimento médio do código como

$$\overline{L} = \sum_{k=0}^{K-1} p_k l_k. \tag{7}$$

O parâmetro $\overline{L}$ representa o número médio de bits por símbolo utilizado no processo de codificação da fonte. Seja $L_{\min}$ o menor valor possível de $\overline{L}$. Define-se eficiência da codificação como

$$\eta = \frac{L_{\min}}{\overline{L}}. \tag{8}$$

Como $\overline{L} \geq L_{\min}$, então $\eta \leq 1$. A codificação é eficiente quando η se aproxima de 1.

O valor de $L_{\min}$ é determinado a partir do primeiro teorema de Shannon (1948):

Dada uma fonte discreta, sem memória[1] *e com entropia* $H(\mathcal{S})$*, o comprimento médio do código é limitado por*

$$\overline{L} \geq H(\mathcal{S}). \tag{9}$$

De acordo com esse teorema, a entropia $H(\mathcal{S})$ representa um limite fundamental no número médio de bits por símbolo necessário para representar uma fonte discreta e sem memória. Desta forma, $L_{\min} = H(\mathcal{S})$ e a eficiência da codificação pode ser reescrita como

$$\eta = \frac{H(\mathcal{S})}{\overline{L}}. \tag{10}$$

Para a demonstração desse teorema consultar as referências (SHANNON, 1948) e (HAYKIN, 2001).

Quando a codificação da fonte permite a recuperação perfeita dos símbolos emitidos fala-se em codificação sem perdas. Existem vários algoritmos para esse tipo de codificação como, por exemplo, códigos de Huffman e Lempel-Ziv. Esses algoritmos geram códigos unicamente decodificáveis. A codificação de Huffman é discutida a seguir.

2.1 Codificação de Huffman

Na codificação de Huffman, palavras-código mais longas são atribuídas às saídas da fonte menos prováveis e palavras mais curtas às mais prováveis. Para fazer isso, inicia-se agrupando as duas saídas da fonte menos prováveis para gerar um novo grupo de saída cuja probabilidade é a soma das probabilidades individuais. O processo é repetido até que reste apenas um grupo de saída gerando dessa forma uma árvore. Começando da raiz dessa árvore e atribuindo 0's e 1's para quaisquer duas ramificações que saem do mesmo nó, gera-se o código. No exemplo seguinte, é mostrado como projetar um código de Huffman.

Exemplo 1 [Código de Huffman] Projete um código de Huffman para uma fonte com alfabeto $\mathcal{X} = x_1,\ x_2,\ \ldots,\ x_9$ e correspondente vetor de probabilidades

$$\boldsymbol{p} = (0.06,\ 0.15,\ 0.1,\ 0.12,\ 0.13,\ 0.07,\ 0.08,\ 0.09,\ 0.2).$$

Encontre o comprimento médio do código resultante e compare com a entropia da fonte.
Solução
Na Figura 2 são apresentadas as etapas do projeto de codificação de Huffman. Primeiramente, deve-se ordenar os elementos do alfabeto em ordem decrescente de probabilidades (Etapa 1) e agrupar os dois menos prováveis (dois últimos elementos). Na etapa seguinte, deve-se novamente ordenar os elementos considerando o grupo formado na etapa anterior, cuja probabilidade é a soma das probabilidades individuais. Deve-se proceder dessa forma até restarem apenas dois grupos (Etapa 8). Nessa etapa, deve-se começar a atribuição das palavras-código. Pode-se atribuir então 1 para o grupo $x_{8,7,2,5,6,1}$ e 0 para o grupo $x_{4,3,9}$. O grupo $x_{4,3,9}$ foi formado na Etapa 6 a partir do grupo $x_{4,3}$ e do elemento x_9. Assim deve-se atribuir, por exemplo, a palavra 01 para o grupo $x_{4,3}$ e a palavra 00 para o elemento x_9 que já pertence ao alfabeto. Deve-se proceder dessa forma para os demais grupos até finalizar a atribuição de códigos para todos os elementos como é mostrado na Figura 3.

[1] Fala-se de fonte sem memória quando os símbolos são estatisticamente independentes.

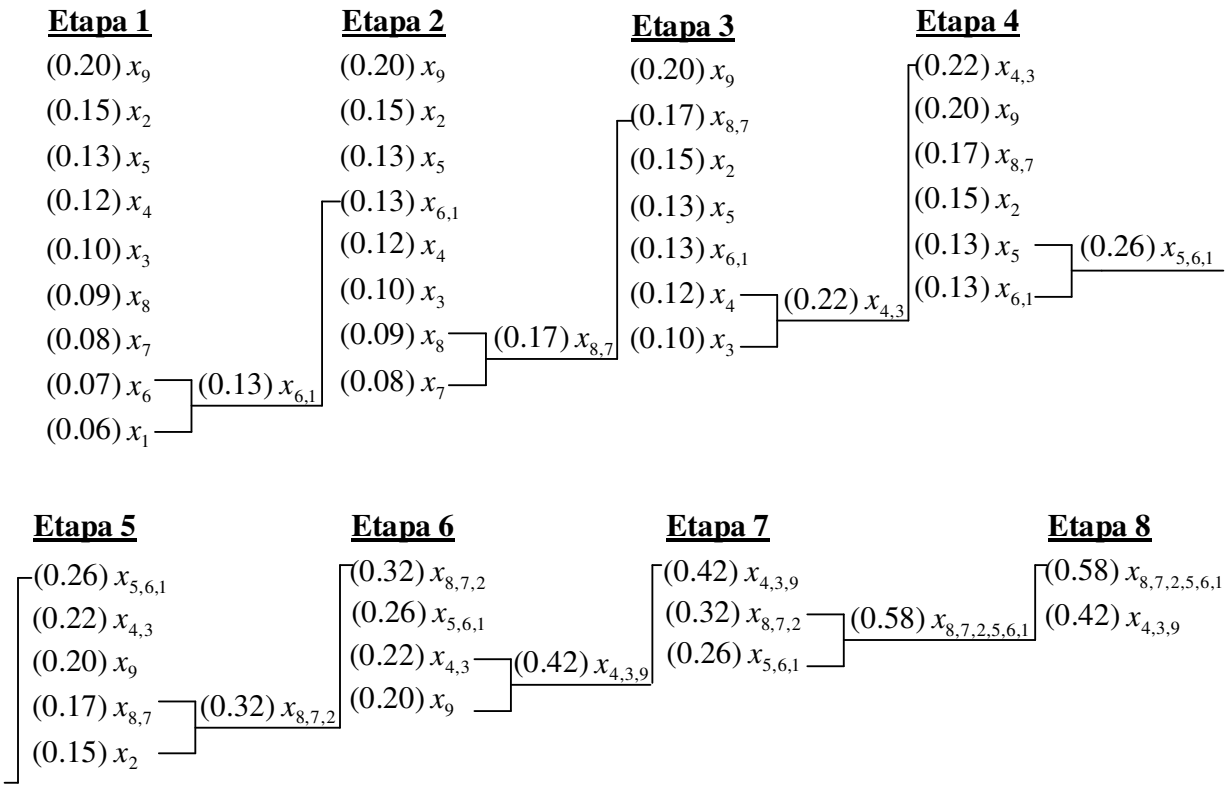

Figura 2: Etapas do projeto da codificação de Huffman

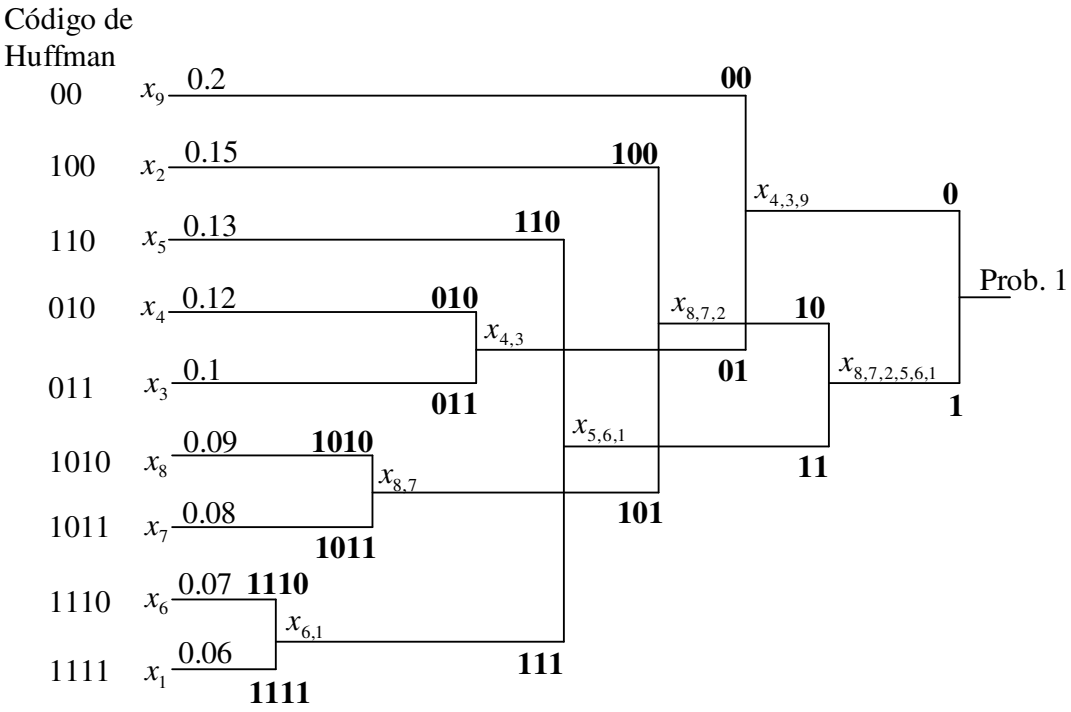

Figura 3: Árvore do código de Huffman

A partir do código de Huffman obtido (Figura 3) pode-se calcular o comprimento médio do código

$$\bar{L} = 2\times 0.2 + 3\times(0.15+0.13+0.12+0.1) + 4\times(0.09+0.08+0.07+0.06) = 3.1 \text{ bits/símbolo.}$$

A entropia da fonte é dada por

$$H(\mathcal{X}) = -\sum_{i=1}^{9} p_i \log p_i = 3.0731 \text{ bits/símbolo.}$$

Como esperado, observa-se que $\bar{L} > H(\mathcal{X})$.

Referências

CARLSON, A. B. *Communication Systems*. 2. ed. New York: McGraw-Hill, 1975.

HAYKIN, S. *Communication Systems*. 4. ed. New York: John Wiley & Sons, 2001.

LATHI, B. P. *Modern Digital and Analog Communication Systems*. 3. ed. New York: Oxford University Press, 1998.

PROAKIS, J. G.; SALEHI, M. *Contemporary Communication Systems using MATLAB*. Pacific Grove: Books/Cole, 2000.

RAMÍREZ, M. A. *Busca de inovações e tratamento de transitórios em codificadores de voz CELP*. 1997. 122p. Tese (Doutorado em Engenharia Elétrica) - Escola Politécnica da Universidade de São Paulo, São Paulo, 1997.

SHANNON, C. E. A mathematical theory of communication. *Bell System Technical Journal*, v. 27, p. 379-423 e p. 623-656, July/Oct. 1948.