

Algoritmos e Estruturas de Dados III

Aula 9.1 – RLE e Métodos Estatísticos

Prof. Felipe Lara



RLE – Run Length Encoding



RLE – Run Length Encoding

Run-Length Encoding (Supressão de Repetições): O método RLE opera reduzindo o tamanho de sequências de símbolos repetidos.

Como: Substitui sequências de caracteres repetidos pelo número de ocorrências seguido do caracter

Exemplo

AAAABBBBAABBBBBBCCCCCCCCCDABCAAABBBB

4A3BAA5B8CDABC3A4B (redução de 32B p/ 18 B)

RLE – Run Length Encoding

Formatos de imagem do tipo bitmap: TIFF (Tag Image File Format), BMP (Microsoft Windows Bitmap), PCX (PC Paintbrush File Format), MacPaint (Macintosh Paint) e TGA (Truevision Graphics Adapter).

Fácil implementação e de execução rápida, mas que não produz taxas de compressão comparáveis com métodos mais complexos, porém mais lentos

Explora a redundância existente entre os pixels de uma imagem (a tendência de que pixels adjacentes possuem valores iguais).

Classificação: Método de compressão simétrico, sem perdas e adaptativo

RLE – Run Length Encoding

Problema: se o texto contiver números, como diferenciar entre os caracteres e o número de repetições?

Usar um caractere especial precedendo o número
&4ABBBAA&5B&8CDABCAAA&4B (32 B p/ 24 B)

Caractere especial não pode ser utilizado no texto

Infelizmente, esse método **não funciona bem para textos** pois normalmente não há muitas repetições de caracteres

RLE – Run Length Encoding

Problema: Expansão de dados. O pior cenário do RLE não é apenas não comprimir, mas sim **aumentar o tamanho do arquivo**.

O Risco da Expansão: Se o dado de entrada não contiver sequências longas (runs), o RLE tem um efeito reverso, **aumentando o tamanho do arquivo**.

Como Ocorre: Cada símbolo original deve ser substituído por **um par** (Contagem + Símbolo), duplicando o número de entradas.

Cenário	Entrada Original	Saída RLE	Impacto
Pior Caso	A B C D E F	1A 1B 1C 1D 1E 1F	Expansão de 100%
	(6 caracteres)	(12 caracteres)	(Dobro do tamanho)

RLE – Run Length Encoding

Adequado para certas sequências binárias (por exemplo: imagens)

- armazena-se para cada linha o número de seqüências de 0's e 1's iniciando-se sempre com 0
- considerando que 5 bits são gastos para cada contagem, o exemplo abaixo reduz o arquivo de 310 para 205 bits

00011111111111111111111111111111	3 28
00111111111111111011111111111111	2 13 1 15
00011111111111110001111111111111	3 11 3 14
00000111111111110001111111111111	5 9 3 14
11000111111111110001111111111111	0 2 3 9 3 14
000000000111111000111111110000000	9 5 3 7 7
11000111111111110001111111111111	0 2 3 9 3 14
00000111111111110001111111111111	5 9 3 14
00111111111111111011111111111111	2 13 1 15
00011111111111111111111111111111	3 28

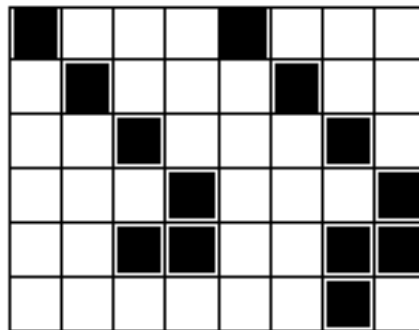
RLE – Run Length Encoding

Simple e rápido tanto para a compressão quanto para a descompressão
Taxa de compressão fortemente dependente da entrada de dados:

- **Imagem preto branco:** contém grande regiões brancas e será comprimida significativamente.
 - compressão boa
- **Imagem com cores chapadas:** tendem a apresentar extensas regiões de mesma cor
 - compressão boa
- **Imagem fotográfica:** muitas e sutis variações de cores/tons tende a não apresentar um boa taxa de compressão
 - tende a ter compressão ruim

RLE – Run Length Encoding

Qual seria o arquivo comprimido no caso do seguinte *bitmap* de 6×8?



Métodos Estatísticos



Métodos Estatísticos

- Utilizam códigos de comprimentos variáveis.
 - Dados na informação original que aparecem com **maior frequência** são representados por **palavras-código menores**
 - Dados de **menor incidência** são representados por **palavras-código maiores**
- Ex: Shannon-Fano / Huffman

Métodos Estatísticos – Entropia

A capacidade de um texto ser comprimido pode ser medida pela **entropia**.

Entropia:

- medida do número médio de símbolos binários (bits) necessários para codificar a saída da fonte.
- na Teoria da Informação (desenvolvida por Claude Shannon), é uma medida fundamental que nos diz o **limite teórico da compressão** que podemos alcançar em um conjunto de dados.
- Usada como uma medida de eficiência para os métodos de compressão.

Métodos Estatísticos – Entropia

Baseada na estimativa (ou cálculo) da frequência de cada símbolo

- Símbolos mais frequentes usarão menos bits
- Símbolos menos frequentes usarão mais bits

Exemplo:

aaaeabbbaaabcaaadacaaabbbaaabcaaaeaaaba

a: 26 vezes

d: 1 vez

b: 8 vezes

e: 2 vezes

c: 3 vezes

Métodos Estatísticos – Entropia

Exemplo:

aaaeabbbaaabcaaadacaaabbbaaabbaaaeaaaba

$$P_a = 26/40 = 0,65$$

$$P_b = 8/40 = 0,2$$

$$P_c = 3/40 = 0,075$$

$$P_d = 1/40 = 0,025$$

$$P_e = 2/40 = 0,05$$

Métodos Estatísticos – Entropia

Exemplo:

aaaeabbbaaabcaaadacaaabbbaaaabbcaaaeaaaba

$$S_x = -\log_2(P_x)$$

- x = Símbolo
- S_x = Entropia de x
- P_x = Probabilidade de x

$$P_a = 26/40 = 0,65$$

$$S_a = \mathbf{0,62} * 26$$

$$P_b = 8/40 = 0,2$$

$$S_b = 2,32 * 8$$

$$P_c = 3/40 = 0,075$$

$$S_c = 3,74 * 3$$

$$P_d = 1/40 = 0,025$$

$$S_d = 5,32 * 1$$

$$P_e = 2/40 = 0,05$$

$$S_e = 4,32 * 2$$

Métodos Estatísticos – Entropia

Exemplo:

aaaeabbbaaabcaaadacaaabbbaaaabbcaaaeaaaba

$$S_x = -\log_2(P_x)$$

- x = Símbolo
- S_x = Entropia de x
- P_x = Probabilidade de x

$$P_a = 26/40 = 0,65 \quad S_a = 0,62 * 26 = 16,10$$

$$P_b = 8/40 = 0,2 \quad S_b = 2,32 * 8 = 18,58$$

$$P_c = 3/40 = 0,075 \quad S_c = 3,74 * 3 = 11,21$$

$$P_d = 1/40 = 0,025 \quad S_d = 5,32 * 1 = 5,32$$

$$P_e = 2/40 = 0,05 \quad S_e = 4,32 * 2 = 8,54$$

59,77

Métodos Estatísticos – Entropia

Exemplo:

aaaebbbaaaabcaaadacaaabbbaaaabbcaaaeeeaba

$$S_x = -\log_2(P_x)$$

- x = Símbolo
- S_x = Entropia de x
- P_x = Probabilidade de x

$$P_a = 26/40 = \mathbf{0,65}$$

$$S_a = 0,62 * 26 = 16,10$$

$$P_b = 8/40 = \mathbf{0,2}$$

$$S_b = 2,32 * 8 = 18,58$$

$$P_c = 3/40 = \mathbf{0,075}$$

$$S_c = 3,74 * 3 = 11,21$$

$$P_d = 1/40 = \mathbf{0,025}$$

$$S_d = 5,32 * 1 = 5,32$$

$$P_e = 2/40 = \mathbf{0,05}$$

$$S_e = 4,32 * 2 = 8,54$$

Probabilidade de
cada símbolo

59,77

Métodos Estatísticos – Entropia

Exemplo:

aaaaebbbaaaabcaaadacaaabbbaaaabbcaaaeeaaaba

$$S_x = -\log_2(P_x)$$

- x = Símbolo
- S_x = Entropia de x
- P_x = Probabilidade de x

Probabilidade de
cada símbolo

$$P_a = 26/40 = \mathbf{0,65}$$

$$P_b = 8/40 = \mathbf{0,2}$$

$$P_c = 3/40 = \mathbf{0,075}$$

$$P_d = 1/40 = \mathbf{0,025}$$

$$P_e = 2/40 = \mathbf{0,05}$$

Entropia de cada
símbolo

$$S_a = \mathbf{0,62} * 26 = 16,10$$

$$S_b = \mathbf{2,32} * 8 = 18,58$$

$$S_c = \mathbf{3,74} * 3 = 11,21$$

$$S_d = \mathbf{5,32} * 1 = 5,32$$

$$S_e = \mathbf{4,32} * 2 = 8,54$$

59,77

Métodos Estatísticos – Entropia

Exemplo:

aaaebbbaaaabcaaadacaaabbbaaaabbcaaaeeaaaba

$$S_x = -\log_2(P_x)$$

- x = Símbolo
- S_x = Entropia de x
- P_x = Probabilidade de x

$$P_a = 26/40 = 0,65$$

$$P_b = 8/40 = 0,2$$

$$P_c = 3/40 = 0,075$$

$$P_d = 1/40 = 0,025$$

$$P_e = 2/40 = 0,05$$

Entropia de cada símbolo

$$S_a = 0,62 * 26 = 16,10$$

$$S_b = 2,32 * 8 = 18,58$$

$$S_c = 3,74 * 3 = 11,21$$

$$S_d = 5,32 * 1 = 5,32$$

$$S_e = 4,32 * 2 = 8,54$$

Probabilidade de cada símbolo

Entropia Total

59,77

Métodos Estatísticos – Entropia

$$S_x = -\log_2(P_x)$$

Símbolo	Probabilidade	Entropia para cada símbolo	Entropia total
U	12/72		
V	18/72		
W	7/72		
X	15/72		
Y	20/72		

Métodos Estatísticos – Entropia

$$S_x = -\log_2(P_x)$$

Símbolo	Probabilidade	Entropia para cada símbolo	Entropia total
U	12/72	2,584963	31,01955
V	18/72	2,000000	36,00000
W	7/72	3,362570	23,53799
X	15/72	2,263034	33,94552
Y	20/72	1,847997	36,95994

Métodos Estatísticos – Entropia

Considere o seguinte texto de 20 caracteres, que será a nossa "Fonte de Informação":

MENSAGEM: CATARATA CACHOEIRA

Responda às seguintes questões:

1. **Cálculo da Probabilidade (P_x):** Calcule a frequência e a probabilidade de cada símbolo na mensagem.
2. **Cálculo da Entropia Individual (S_x):** Calcule a Entropia Individual (S_x) de cada símbolo, usando a fórmula de Shannon: $S_x = -\log_2(P_x)$.
3. **Entropia Total da Mensagem:** Calcule a **Entropia Total** (S_{Total}) da mensagem.
4. **Análise da Codificação Fixa:** a) Qual é o **número mínimo de bits** necessário para representar **cada símbolo** se usarmos um código de **tamanho fixo**? b) Qual seria o **tamanho total** da mensagem usando essa codificação de tamanho fixo?
5. **Conclusão sobre a Compressão:** Qual a **taxa de compressão máxima** esperada, comparando o tamanho da Entropia Total (item 3) com o tamanho da Codificação Fixa (item 5b)?

Shannon Fano



PUC Minas

Shannon-Fano

- Apresentado por C. E. Shannon e por R. M. Fano em 1949.
- O objetivo deste método é associar **códigos menores a símbolos mais prováveis** e **códigos maiores aos menos prováveis**.
- A codificação parte da construção de uma árvore ponderada considerando o **peso de cada símbolo**, ou seja, a probabilidade de ocorrência do mesmo em um texto.

Shannon-Fano - Algoritmo

- 1) Criar uma lista de **probabilidades ou contagens de frequência** para o determinado **conjunto de símbolos** de forma que a frequência relativa de ocorrência de cada símbolo seja conhecida.
- 2) Classificar a lista de símbolos em ordem **decrescente de probabilidade**, os mais prováveis à esquerda e os menos prováveis à direita.
- 3) Dividir a **lista em duas partes**, com a probabilidade **total** de ambas as partes sendo a mais próxima possível.
- 4) Atribuir o valor 0 à parte esquerda e 1 à parte direita.
- 5) Repetir os passos 3 e 4 para cada parte, até que todos os símbolos sejam divididos em subgrupos individuais.

- <https://wordcounter.net/character-count>

Shannon-Fano

Por exemplo, seja a seguinte lista $L = \{s1, s2, s3, s4, s5\}$ com respectivos pesos $P = \{0,3; 0,2; 0,2; 0,2; 0,1\}$ obtidos a partir da seguinte sequência de símbolos.

- [illegible]

Shannon-Fano

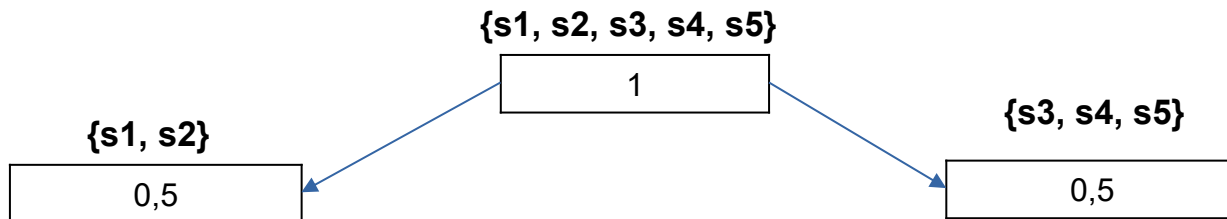
Por exemplo, seja a seguinte lista $L = \{s1, s2, s3, s4, s5\}$
com respectivos pesos $P = \{0,3; 0,2; 0,2; 0,2; 0,1\}$

$\{s1, s2, s3, s4, s5\}$

1

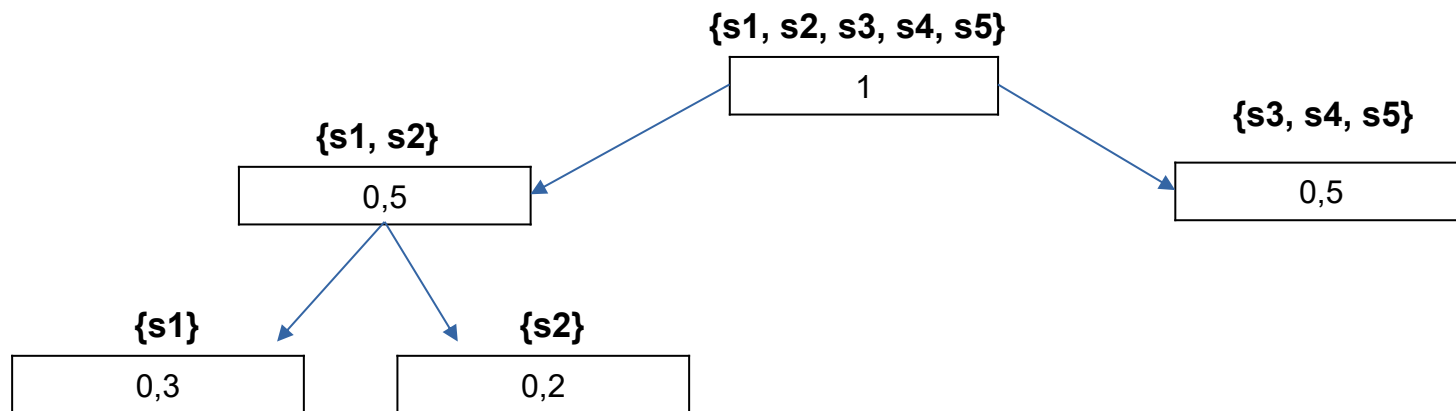
Shannon-Fano

Por exemplo, seja a seguinte lista $L = \{s1, s2, s3, s4, s5\}$
com respectivos pesos $P = \{0,3; 0,2; 0,2; 0,2; 0,1\}$



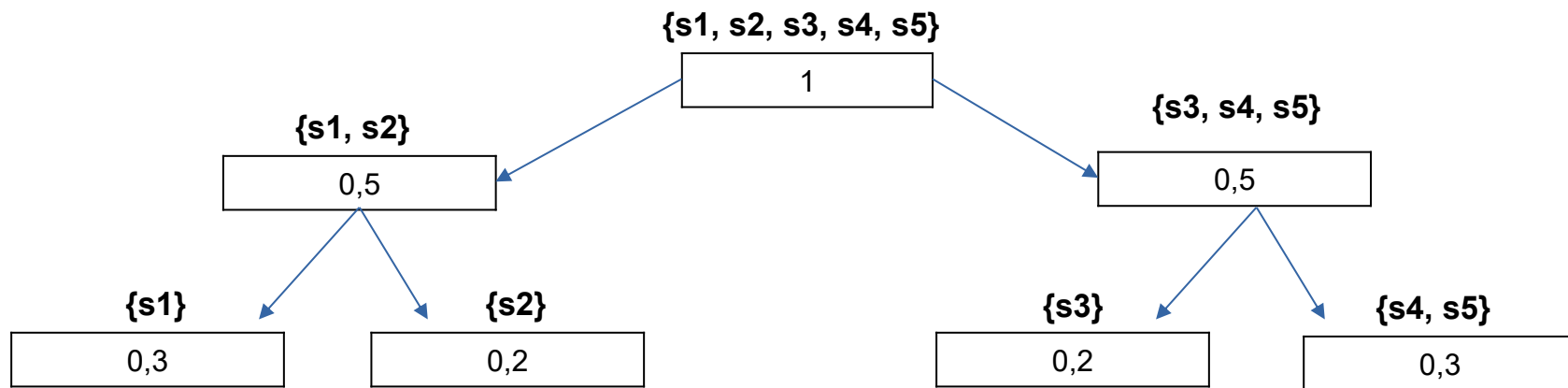
Shannon-Fano

Por exemplo, seja a seguinte lista $L = \{s1, s2, s3, s4, s5\}$
com respectivos pesos $P = \{0,3; 0,2; 0,2; 0,2; 0,1\}$



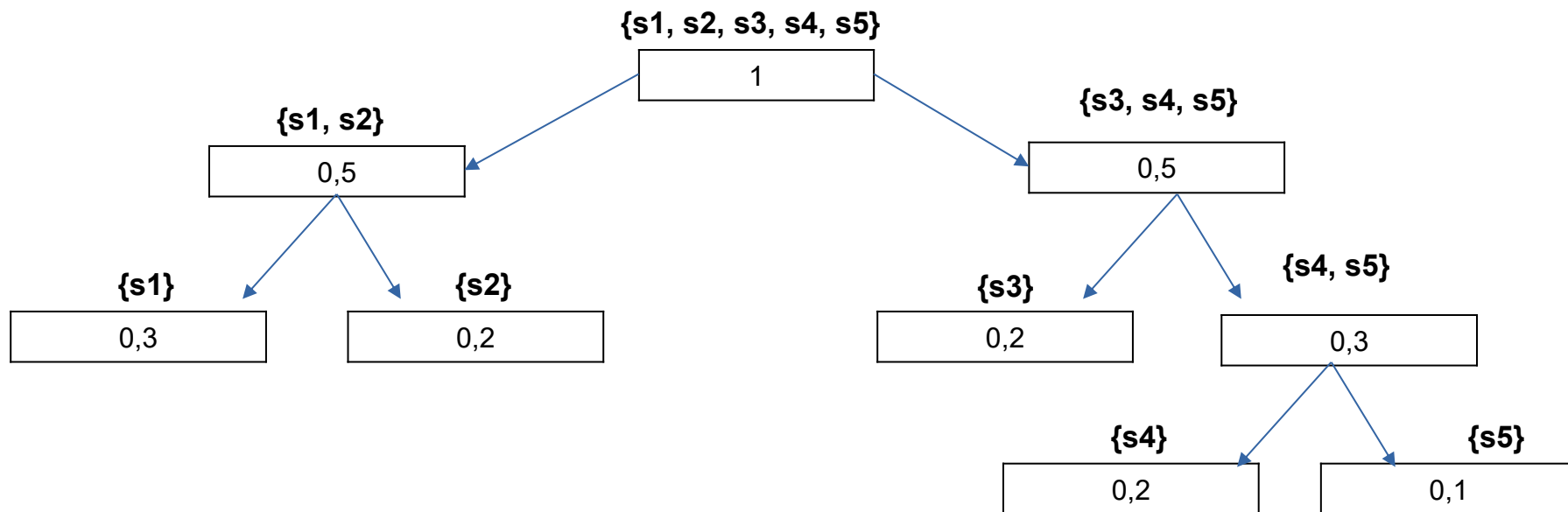
Shannon-Fano

Por exemplo, seja a seguinte lista $L = \{s1, s2, s3, s4, s5\}$
com respectivos pesos $P = \{0,3; 0,2; 0,2; 0,2; 0,1\}$



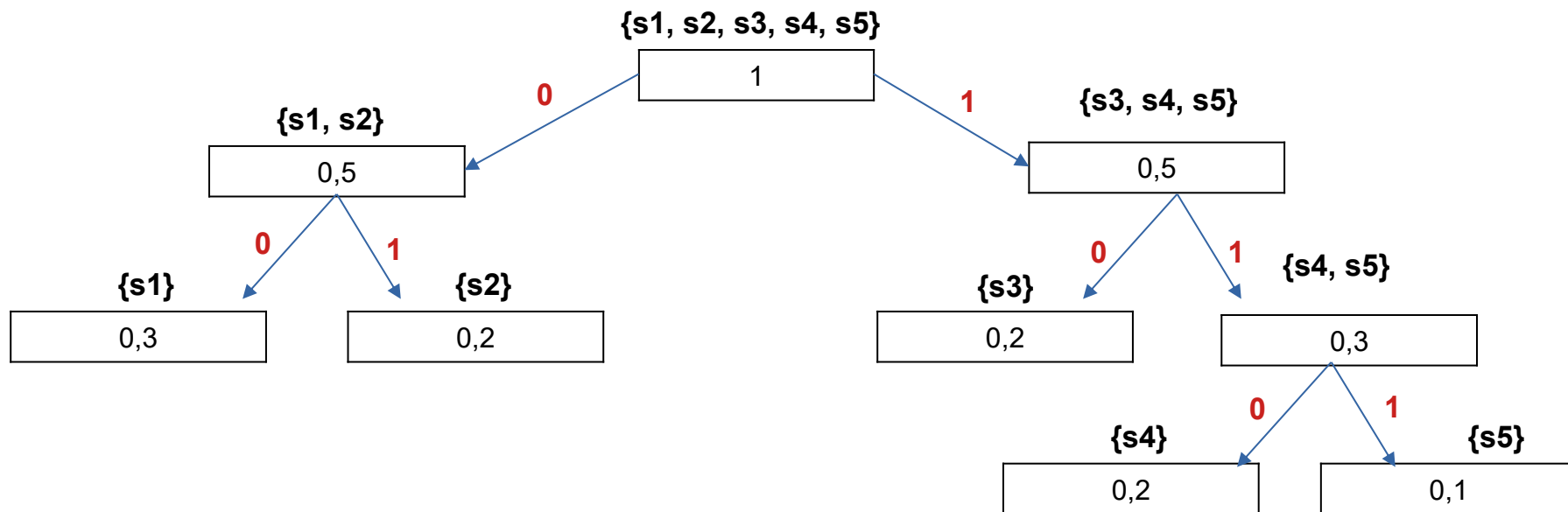
Shannon-Fano

Por exemplo, seja a seguinte lista $L = \{s1, s2, s3, s4, s5\}$
com respectivos pesos $P = \{0,3; 0,2; 0,2; 0,2; 0,1\}$



Shannon-Fano

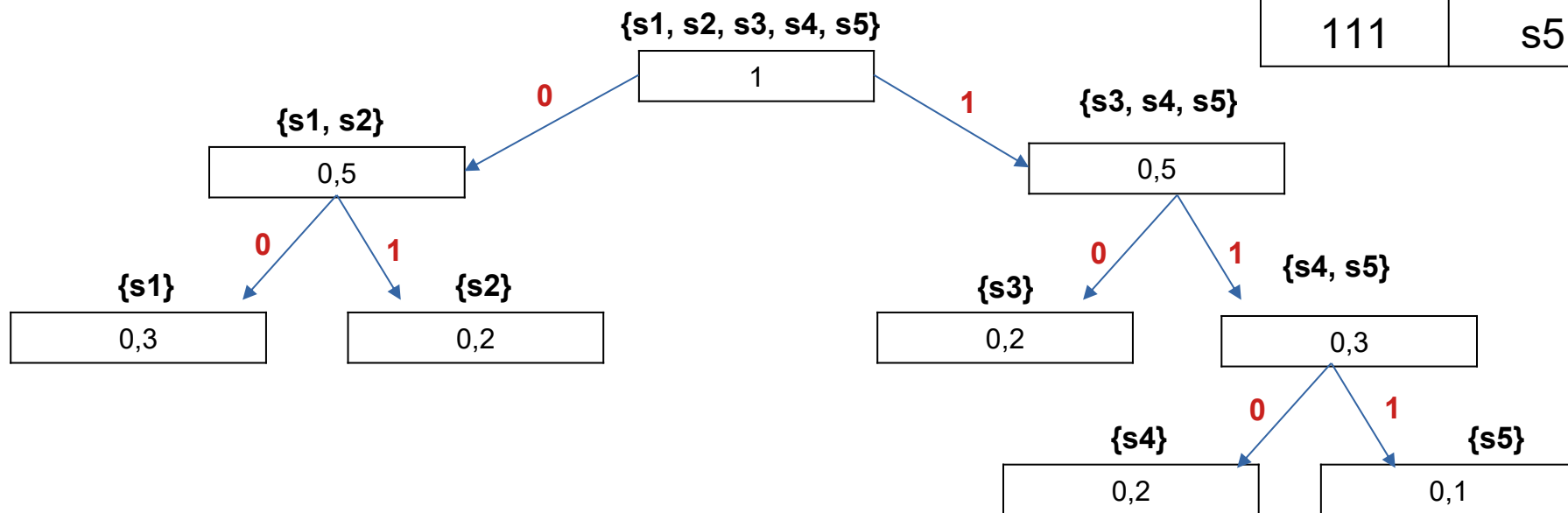
Por exemplo, seja a seguinte lista $L = \{s1, s2, s3, s4, s5\}$
com respectivos pesos $P = \{0,3; 0,2; 0,2; 0,2; 0,1\}$



Shannon-Fano

Por exemplo, seja a seguinte lista $L = \{s1, s2, s3, s4, s5\}$
com respectivos pesos $P = \{0,3; 0,2; 0,2; 0,2; 0,1\}$

Código	Símbolo
00	s1
01	s2
10	s3
110	s4
111	s5



Shannon-Fano

- Sem perdas
- Eficiente e prático, mas gera resultados sub-ótimos
- Sua aplicação prática é quase nula em relação ao método de Compressão Huffman
- Huffman: constrói a árvore binária de forma bottom-up!

Huffman



PUC Minas

Huffman

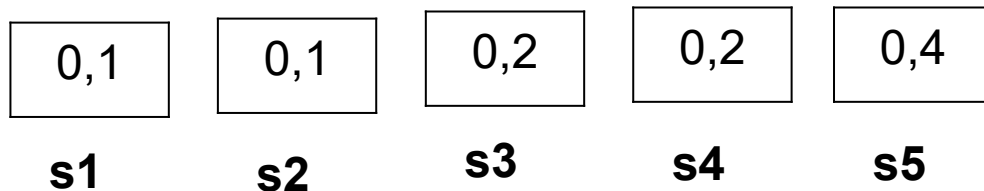
- Proposto por Huffman em 1952
- Método com o objetivo de obter a redundância mínima desejada do texto comprimido.
- Construção de uma árvore de menor altura ponderada.
- Como em Shannon-Fano, considera-se a existência de um alfabeto fonte, onde cada símbolo tem seu respectivo peso.
- Como o objetivo é criar um código de prefixo mínimo, associamos códigos menores a símbolos mais prováveis e códigos maiores a símbolos menos prováveis.

Huffman - Algoritmo

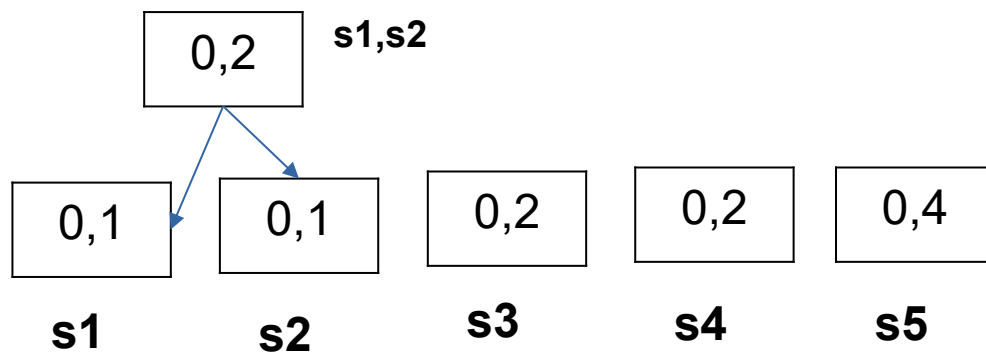
- 1) Considere uma floresta em que cada árvore tenha sua raiz associada a um símbolo do alfabeto com seu respectivo peso;
- 2) Remova quaisquer duas árvores cujas raízes tenham menor peso. Acrescente uma nova árvore que tenha uma raiz cujos filhos sejam árvores anteriores e cujo peso seja a soma dos pesos das raízes dessas árvores;
- 3) Repita o passo anterior até que exista somente uma árvore na floresta.

Huffman - Exemplo

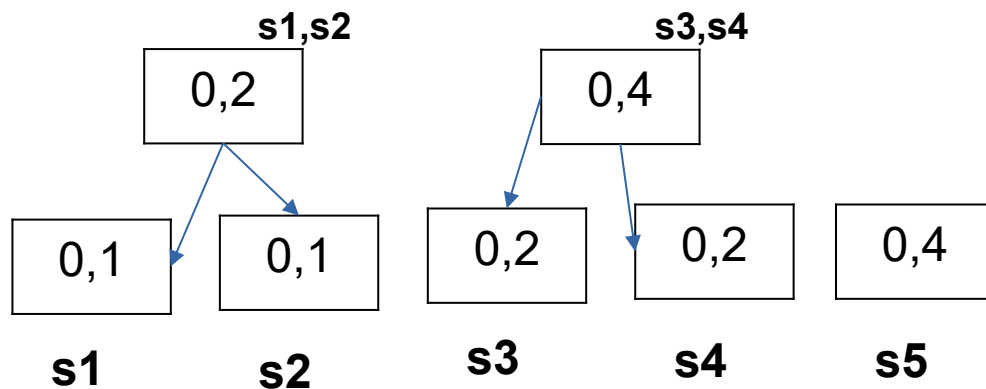
Por exemplo, seja a seguinte lista $L = \{s1, s2, s3, s4, s5\}$
com respectivos pesos $P = \{0,1; 0,1; 0,2; 0,2; 0,4\}$
temos a seguinte floresta inicial



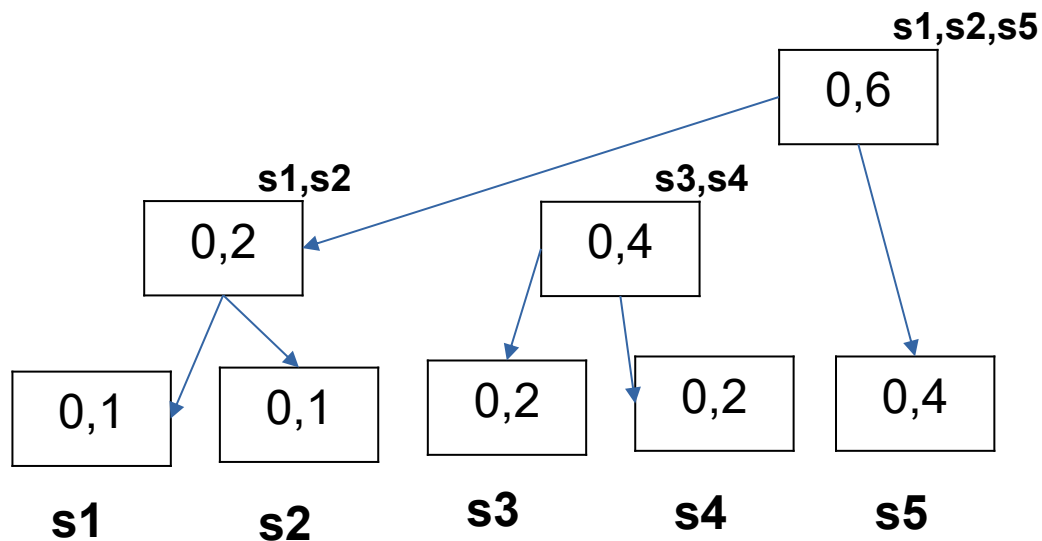
Huffman - Exemplo



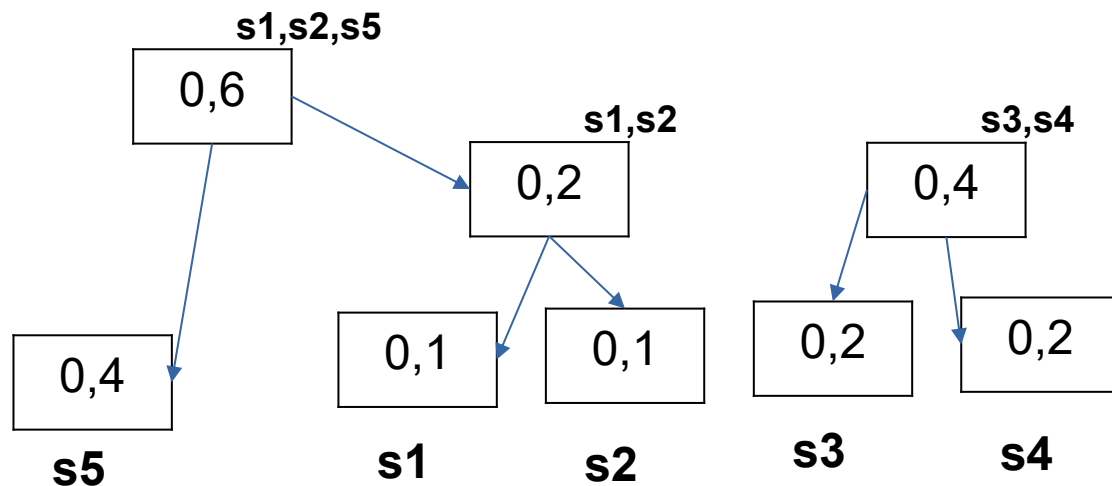
Huffman - Exemplo



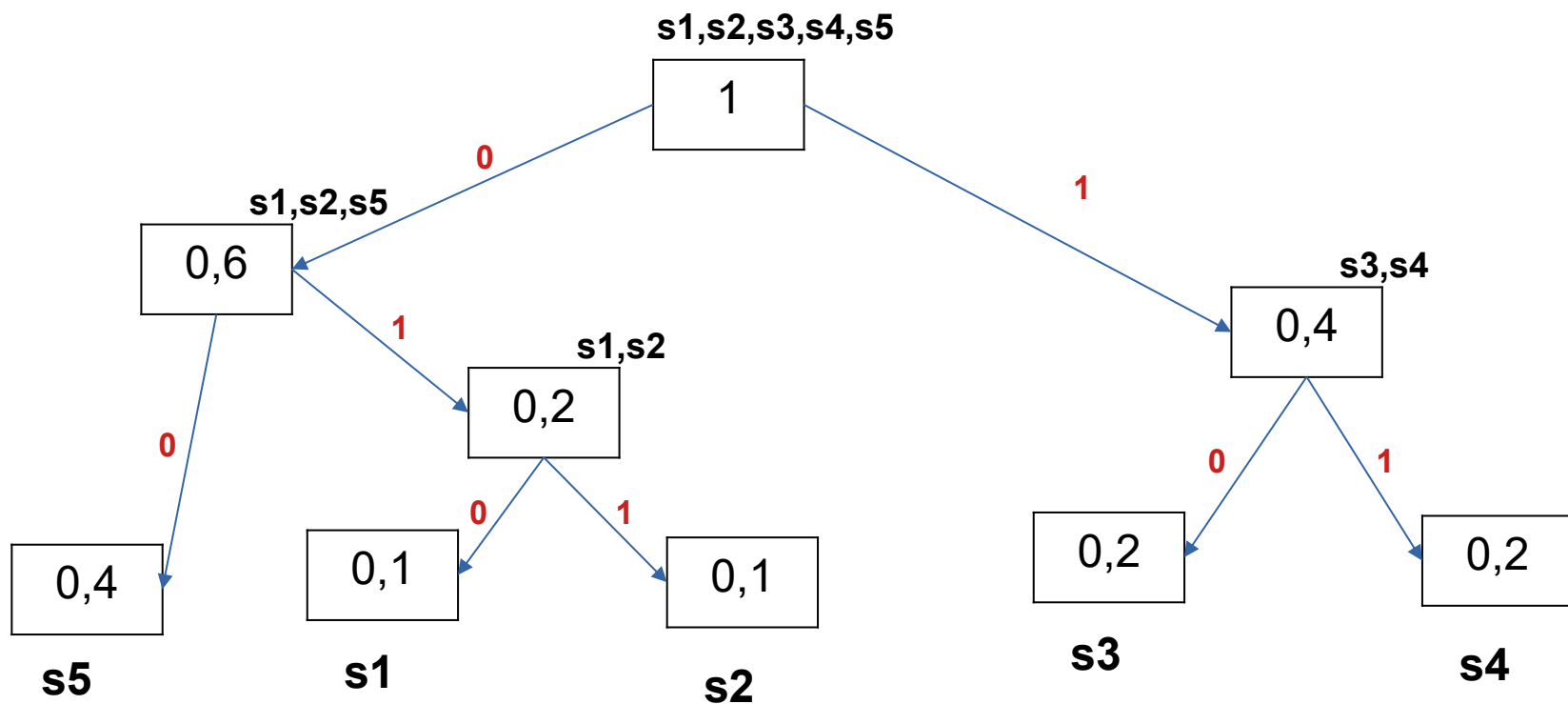
Huffman - Exemplo



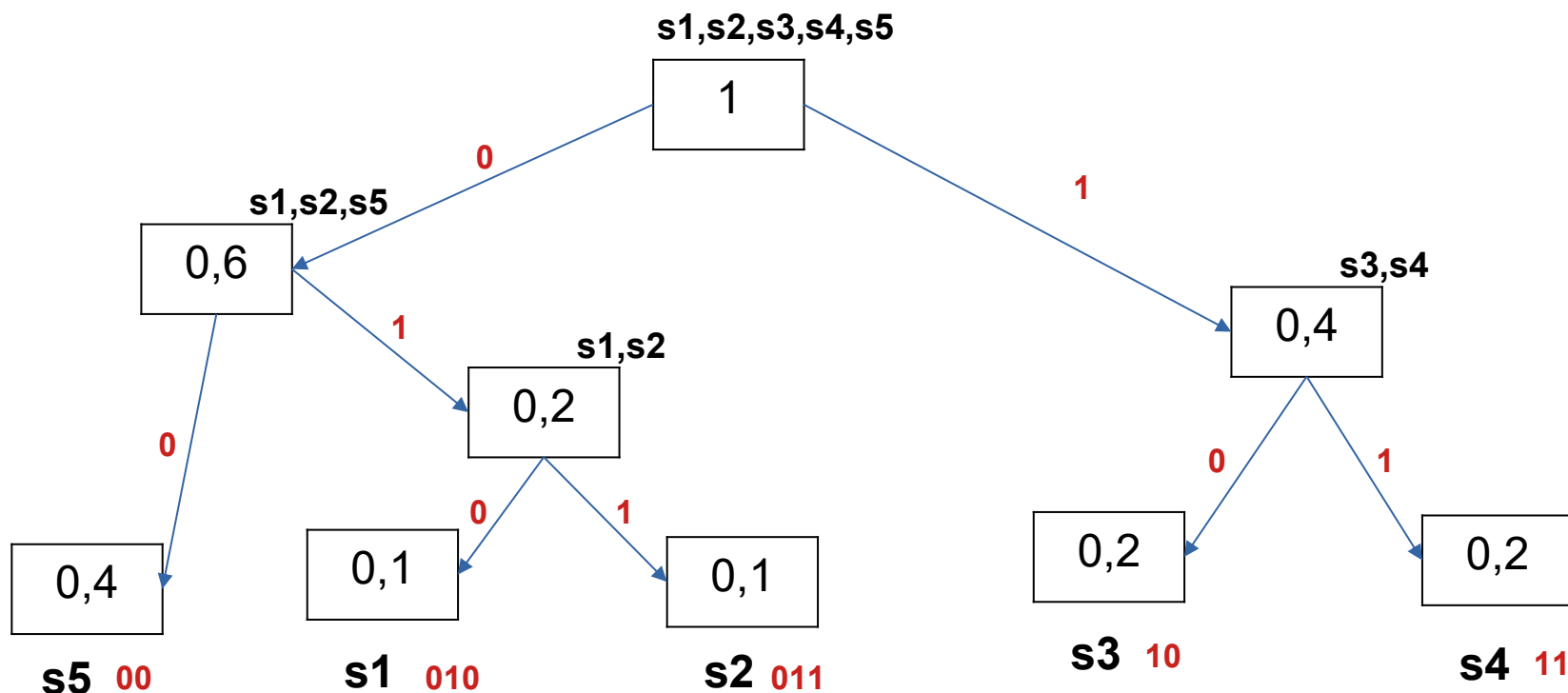
Huffman - Exemplo



Huffman - Exemplo

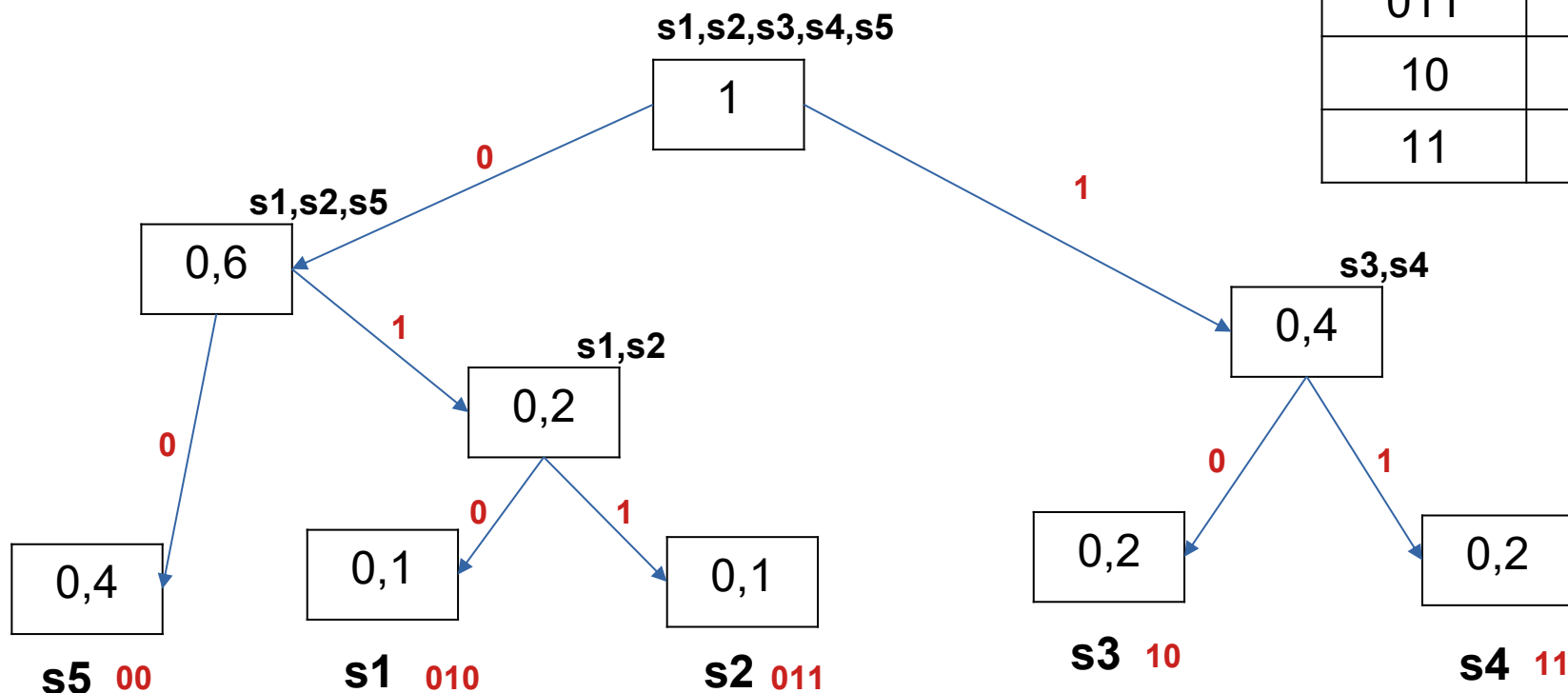


Huffman - Exemplo



Huffman - Exemplo

Código	Símbolo
00	s5
010	s1
011	s2
10	s3
11	s4



Huffman - Exemplo

s1s1s5s4s3

=

010010001110

Código	Símbolo
00	s5
010	s1
011	s2
10	s3
11	s4

Huffman

Sem perdas

Construção da a árvore binária de forma bottom-up!

A árvore não é única, mas garante códigos de redundância mínima.

Necessita de duas leituras sobre o texto fonte = deficiente em alguns casos, como por exemplo na transmissão de dados

Exercício

A ARANHA ARRANHA A RÃ

Calcule entropia, Shannon-Fano e Huffman.