Projeto
Introdução à Teoria da Informação
Prof. Derzu Omaia

Projeto 2022.1

Parte 1:
Implemente um compressor e descompressor utilizando o algoritmo LZW ou utilizando o PPM-C (visto em sala de aula) com o codificador aritmético. Considere que as mensagens são geradas por fontes com alfabeto A = {0, 1, ..., 255}. Teste o compressor/descompressor com um arquivo binário de vídeo e com um corpus de texto em português de 16MB.

LZW:

Teste o compressor com diferentes tamanhos máximos de dicionário. O tamanho máximo do dicionário deve ser 2^K, onde K deve variar entre 9 e 16. O K indica a quantidade de bits com que cadas índice do dicionário deve ser armazenado.

Exemplo: K=9bits, tamanho do dicionário: 2^9=512, K=10bits tamanho do dicionário $2^{\wedge 10}$=1024.

No relatório apresente as curvas de RC x K e de Tempo de Processamento x K, para K = 9, 10, 11, 12, 13, 14, 15, 16 bits. A variação do K deve ser realizada no eixo x do gráfico (horizontal). Indique também a quantidade total de índices presentes na mensagem final para cada K.

PPM-C + Aritmético:

O contexto deve ter tamanho máximo K (parâmetro). O modelo PPM-C alimentará um codificador aritmético. Utilize o mecanismo de exclusão quando necessário. No relatório apresente as curvas de RC x K e de Tempo de Processamento x K, para K = 0, 1, 2, 3, 4, 5, 6, 7, 8. A variação do K deve ser realizada no eixo x do gráfico (horizontal). Não é necessário implementar o codificador aritmético, utilize algum já existente.

Observações:

- Os símbolos do arquivo de teste devem ser lidos no modo binário (números) e não no modo texto (caracteres/strings).
- Dicionário com símbolos numéricos, não caracteres. Utilize arrays inteiros e não strings.
- O codificador deve receber como entrada um arquivo e gerar como saída o arquivo codificado. O decodificador deve receber como entrada o arquivo codificado e gerar como saída o arquivo decodificado, este deve ser igual ao original.

- A execução dos experimentos é demorada, evite fazer os experimentos na véspera da entrega pois não dará tempo.

Pontos a serem avaliados no relatório:

1. Desenvolveu em qual linguagem? Fez utilizando PPMC ou LZW? Utilizou alguma biblioteca como base?

2. A compressão e descompressão funcionou para os 2 arquivos de testes?
    a. O arquivo de vídeo continua sendo "tocável" após a descompressão?
    b. No arquivo de texto, teve algum problema com os caracteres acentuados?
    c. Se não funcionou perfeitamente, qual o problema que ocorreu?
    d. Utilize o programa diff ou FC para comparar se o arquivo original é igual ao arquivo descomprimido. Informe se a comparação indicou que os arquivo são iguais.

3. Conseguiu fazer com todos os valores de Ks solicitados (9 a 16 no LZW ou 0 a 8 no PPM)?

4. Conseguiu salvar a quantidade exata de K bits no arquivo?
    a. Usou qual técnica? Conversão de bits para String? Bitstream? Salvou todos os índices em 2 bytes?

5. Apresentou a curva RCxK e Tempo x K para os 2 arquivos de testes?
    a. Como calculou a RC?

    RC = tamArqOriginal / tamArqComprimido
    RC = tamArqOriginal / ((totalIndices*K)/8)

6. Dicionário ficou estático depois que atingiu o tamanho máximo? Em que Ks ele atingiu o tamanho máximo?

Parte 2:

Implementar um reconhecedor de padrões baseado em PPM ou LZW. Utilize um banco de dados previamente rotulado, na etapa de treinamento gere um modelo PPM (árvore) ou LZW (dicionário) para cada categoria do banco de dados selecionado.

Organize o banco de dados em amostras de treino e classificação utilizando a técnica de validação cruzada. Isto é, para cada categoria do banco de dados, selecione todas as amostras - 1 para treinamento e 1 amostra para classificação, a seleção dessas amostras deve ser aleatória. Para classificação utilize o algoritmo dos K-Vizinhos mais Próximos (K-NN, *k-nearest neighbors*), com k=1, utilizando como métrica de distância o tamanho do arquivo comprimido. No caso do LZW, outra possibilidade de métrica de distância é a quantidade de índices utilizado pelo LZW, antes da serialização para bytes. A Figura 1 apresenta esse processo de validação cruzada para um banco de dados de texturas de imagens.

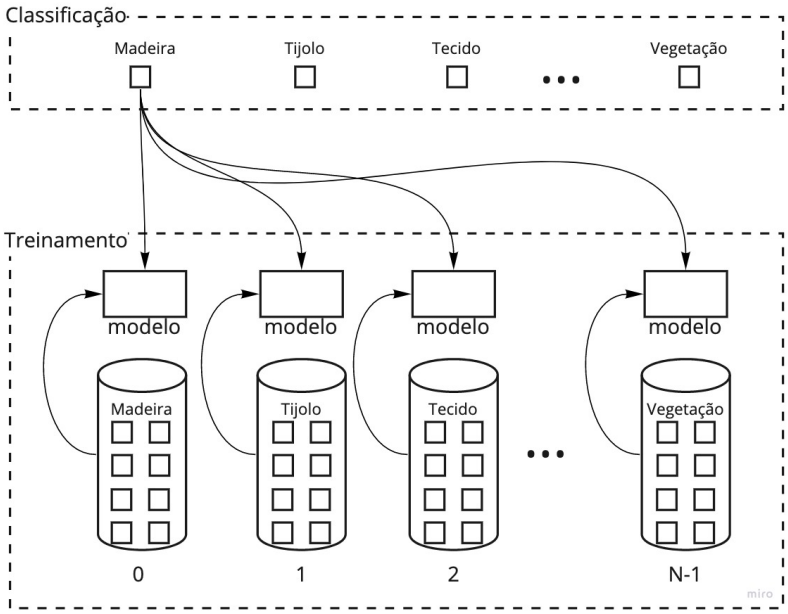

Figura 1 – Esquema de validação cruzada.

O treinamento consiste em gerar o dicionário de cada categoria do banco de dados. A classificação/testes consiste em comprimir 1 amostra (que não foi utilizada na geração do modelo) em todos os modelos/dicionários. Durante a compressão da amostra de teste o dicionário deve permanecer estático. A amostra de teste deve ser atribuída ao modelo que proporcionou a melhor compressão.

No caso do PPM, teste com contexto variável, K = 0, 1, 2, 3, 4, 5, 6, 7, 8. Para o LZW, teste com dicionários de tamanho 2^K, com K= 9, 10, 11, 12, 13, 14, 15, 16. Faça um relatório e neste apresente as curvas de Taxa de acerto x K, e de Tempo de Processamento x K, para.

Sugestões de bancos de dados para serem utilizados:

1. Iris:
    a. Banco de dados: *Iris Database Palacký University (http://phoenix.inf.upol.cz/iris/download/)*
    b. Utilizar as 40 das 64 pessoas disponíveis no banco. Cada pessoa possui 6 fotos de sua íris.

2. Face
    a. Banco de dados: *ORL Database of Faces*. (https://www.dropbox.com/s/mnhfhb1i51loknk/orl_faces.zip?dl=0)
    b. Utilizar as 40 pessoas disponíveis no banco. Cada pessoa possui 10 fotos de sua face.

3. Instrumentos Musicais
    a. *IRMAS Data Base.* (https://www.dropbox.com/s/gebspyfkse1bkju/IRMAS-TrainingData_red.zip?dl=0)

b. 10 instrumentos musicais disponíveis no banco. Cada instrumento possui aproximadamente 100 amostras de áudio.

4. Texturas
    a. *Brodatz Data Base* (https://www.dropbox.com/s/vvzg6xbbcfodn1a/brodatz.zip)
    b. 100 texturas de imagem disponíveis no banco. Cada textura possui apenas 1 amostra. Utilizar as imagens combinadas para classificação.

5. Voz
    a. Escolher o banco.
6. Autores de livros
    a. Escolher/fazer o banco.

Observações:

1. Utilizar as imagens em escala de cinza.
2. Gerar um relatório com o resultado dos experimentos.
3. Grupos de até 3 pessoas.

Questões a serem levantadas no relatório:

1 - Utilizou qual banco de dados?
2 - Se foi um banco de dados de imagem, descartou o cabeçalho do arquivo, e usou apenas a sequência pixels?
3 - Utilizou qual linguagem de programação?
4 - Gerou o banco de treinamento e a arquivo de teste de forma aleatória? Ficaram quantos arquivos para treino e quantos para classificação?
5 - Testou com todos os K (9 a 16 no LZW ou 0 a 8 no PPM)?
6 - Dicionário ficou estático durante a classificação?
7 - Apresentou as curvas de Taxa de acerto x K, e de Tempo de Processamento x K?. Lembrando que K deve ser no eixo X.
8 - Usou qual métrica de distância? Tamanho do arquivo ou quantidade de índices?

Entrega:

Parte 1: 17/11/2022. Cada dia de atraso reduz em 10% a nota máxima.

Parte 2: 08/12/2022. Cada dia de atraso reduz em 10% a nota máxima.