Compressão estatística de textos

Foi visto até aqui um método específico de compressão de texto baseado em dados estatísticos e utilizado na codificação Huffman. Nele, a modelagem estatística do texto é realizada atribuindo probabilidades aos caracteres de acordo com suas respectivas frequências de aparição no texto e, em seguida, a codificação é feita atribuindo códigos menores aos caracteres mais prováveis e códigos maiores aos menos prováveis.

Porém, este modelo estatístico não é o único utilizado. Salomon (2007, p. 139) explica, também, a abordagem estatística baseada no contexto, a qual atribui probabilidades a dados futuros fundamentadas em dados que já foram processados. No caso de textos, então, o contexto de um dado caractere são os caracteres já lidos previamente e a atribuição de probabilidades aos caracteres que não foram compactados ainda depende do contexto desses caracteres.

Modelagens estatísticas baseadas em contexto serão explicadas na seção sobre Cadeias de Markov baseadas em contexto.

Cadeias de Markov

Segundo Ross (2010, p. 39), na matemática é possível estudar eventos de acaso, que são experimentos cujos resultados não são absolutamente previsíveis. Dentro dessa classe de eventos há aqueles que são

independentes, isto é, o resultado de um evento independente decorre exclusivamente do acontecimento do próprio evento, sem a influência do resultado de outros eventos. Porém, conforme Grinstead e Snell (2006, p. 405), há também eventos cujos resultados conhecidos podem influenciar os resultados de experimentos futuros. Processos dessa natureza, em que o resultado de um evento influencia o resultado de um evento futuro são conhecidos como Cadeias de Markov.

As cadeias de Markov são capazes de modelar processos em que há mudanças de estados, partindo de um estado inicial e passando por uma sequência de outros estados sucessivos. Neste contexto, estado é a situação em que um sistema ou processo se encontra em determinado momento, enquanto transição de estados é a passagem de um estado a outro e ocorre com uma determinada probabilidade.

Uma cadeia de Markov, segundo Weber (2011, p. IV), é um processo de transições aleatórias de estados cujo estado futuro depende apenas do estado atual e que não possui memória de como este foi alcançado. Matematicamente uma cadeia de Markov é descrita por um conjunto de estados $\boldsymbol{S} = \{\boldsymbol{s_1}, \boldsymbol{s_2}, \dots, \boldsymbol{s_n}\}$ e a transição de estado de $\boldsymbol{s_i}$ para $\boldsymbol{s_j}$ ocorre com uma probabilidade $\boldsymbol{p_{ij}}$, chamada probabilidade de transição.

Dadas essas informações, é possível estabelecer a probabilidade de, a partir de um estado inicial $\boldsymbol{s_i}$, um processo evoluir para um estado final $\boldsymbol{s_j}$

independentemente de quantas transições de estado ocorram. Por exemplo, supondo que na cidade de Santo André, quando um dia está ensolarado, o dia seguinte também estará ensolarado com probabilidade de 0,5, ou estará chuvoso com probabilidade de 0,2 ou nublado com probabilidade de 0,3. Porém, se o dia está chuvoso, o dia seguinte terá probabilidade 0,4 de estar ensolarado, 04 de estar chuvoso e 0,2 de estar nublado. Por fim, a probabilidade de um dia estar ensolarado após um dia chuvoso é de 0,2, de 0,5 para estar nublado e de 0,25 para continuar chuvoso. Dadas essas informações, podemos calcular a probabilidade de, sendo hoje um dia ensolarado, depois de amanhã ser um dia nublado. Para isso, há três casos possíveis:

I) Amanhã estará ensolarado e, depois de amanhã, nublado;

II) Amanhã estará chuvoso e, depois de amanhã, nublado;

III) Amanhã estará nublado e, depois de amanhã, nublado;

Para representar todas as transições entre os estados de uma cadeia de Markov, podemos montar uma matriz quadrada cuja ordem seja a cardinalidade do conjunto de estados $\boldsymbol{S}$, isto é, a quantidade de estados da cadeia, e o conteúdo dessa matriz quadrada sejam todas as probabilidades de transição entre os estados. Cada linha e cada coluna dessa matriz caracterizam um estado da cadeia em que as linhas representam o estado de partida e as colunas o estado de chegada das transições. À essa matriz, que armazena as probabilidade de transição entre estados - estes representados

por linhas e colunas - dá-se o nome de matriz de transição ou matriz de probabilidades de transição, representada por $\boldsymbol{T}$.

Com isso, podemos então montar a matriz de transição do exemplo anterior, cujo conjunto de estados $\boldsymbol{S}$ é dado por $\boldsymbol{S} = \{ensolarado\ (E),\ nublado\ (N),\ chuvoso\ (C)\}$. Como este conjunto tem cardinalidade $|\boldsymbol{S}| = 3$, então a matriz de transição $\boldsymbol{T}$ será de ordem 3 e cada posição da matriz armazenará uma probabilidade de transição, conforme imagem abaixo:

$$\boldsymbol{T} = \begin{array}{c} \\ \text{De} \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \multicolumn{3}{c}{\text{Para}} \\ \text{E} & \text{C} & \text{N} \end{array} \\ \begin{bmatrix} 0{,}50 & 0{,}20 & 0{,}30 \\ 0{,}40 & 0{,}40 & 0{,}20 \\ 0{,}25 & 0{,}50 & 0{,}25 \end{bmatrix} \end{array} \begin{array}{l} \\ \text{E} \\ \text{C} \\ \text{N} \end{array}$$

Agora, com esta matriz de transição construída é mais simples calcular a probabilidade de ocorrer uma sequência qualquer de estados. Cada probabilidade de transição é dada por p_{ij}, em que i representa o estado atual ($i - ésim$ linha de $\boldsymbol{T}$) e j o estado futuro ($j - ésima$ coluna de $\boldsymbol{T}$) e a probabilidade de uma transição de estados ocorrer em n etapas é dada por $p_{ij}^{(n)}$ da matriz de transição $\boldsymbol{T^n}$ ($n - ésima$ potência da matriz $\boldsymbol{T}$). Por exemplo, um estado i tem probabilidade $p_{ij}^{(1)} = p_{ij}$ de mudar para o estado j em apenas uma etapa. Da mesma forma, se fossem necessárias duas etapas para o estado ir de i a j, então a probabilidade seria $p_{ij}^{(2)}$, e assim sucessivamente.

Retomando o exemplo da transição do tempo na cidade de Santo André, a seguir vamos simular as transições entre estados em 2, 3, 4 e 5 etapas, isto é, calcular $\boldsymbol{T^2}$, $\boldsymbol{T^3}$, $\boldsymbol{T^4}$ e $\boldsymbol{T^5}$:

$$\boldsymbol{T^2} = \begin{array}{c} \\ \text{De} \\ \\ \\ \end{array} \begin{array}{c} \text{Para} \\ \begin{array}{ccc} \text{E} & \text{C} & \text{N} \end{array} \\ \begin{array}{c} \text{E} \\ \text{C} \\ \text{N} \end{array} \begin{bmatrix} 0{,}405 & 0{,}330 & 0{,}265 \\ 0{,}410 & 0{,}340 & 0{,}250 \\ 0{,}388 & 0{,}375 & 0{,}238 \end{bmatrix} \end{array}$$

$$\boldsymbol{T^3} = \begin{array}{c} \\ \text{De} \\ \\ \\ \end{array} \begin{array}{c} \text{Para} \\ \begin{array}{ccc} \text{E} & \text{C} & \text{N} \end{array} \\ \begin{array}{c} \text{E} \\ \text{C} \\ \text{N} \end{array} \begin{bmatrix} 0{,}401 & 0{,}346 & 0{,}254 \\ 0{,}404 & 0{,}343 & 0{,}254 \\ 0{,}403 & 0{,}346 & 0{,}251 \end{bmatrix} \end{array}$$

$$\boldsymbol{T^4} = \begin{array}{c} \\ \text{De} \\ \\ \\ \end{array} \begin{array}{c} \text{Para} \\ \begin{array}{ccc} \text{E} & \text{C} & \text{N} \end{array} \\ \begin{array}{c} \text{E} \\ \text{C} \\ \text{N} \end{array} \begin{bmatrix} 0{,}402 & 0{,}345 & 0{,}253 \\ 0{,}402 & 0{,}345 & 0{,}253 \\ 0{,}403 & 0{,}344 & 0{,}253 \end{bmatrix} \end{array}$$

$$\boldsymbol{T^5} = \begin{array}{c} \\ \text{De} \\ \\ \\ \end{array} \begin{array}{c} \text{Para} \\ \begin{array}{ccc} \text{E} & \text{C} & \text{N} \end{array} \\ \begin{array}{c} \text{E} \\ \text{C} \\ \text{N} \end{array} \begin{bmatrix} 0{,}402 & 0{,}345 & 0{,}253 \\ 0{,}402 & 0{,}345 & 0{,}253 \\ 0{,}402 & 0{,}345 & 0{,}253 \end{bmatrix} \end{array}$$

Por exemplo, se quisermos saber a probabilidade de um dia estar nublado 5 dias após um dia chuvoso basta obter o valor da probabilidade correspondente na matriz $\boldsymbol{T^5}$**,** que é $p_{23}^{(5)} = 0{,}253$.

As matrizes acima, então, mostram o comportamento da cadeia de Markov a longo prazo, ainda que o termo “longo prazo” dependa da cadeia em particular. É possível perceber que a partir da 5ª etapa da cadeia a matriz de probabilidade converge para colunas de mesmas probabilidades, isto é, as colunas 1, 2 e 3 possuem probabilidades iguais a .402, .345 e .253, respectivamente e independentemente de qual tenha sido o estado inicial. Essa classe de cadeias de Markov que a longo prazo convergem suas probabilidades finais a um mesmo valor, tornando a análise independente do estado inicial, são chamadas de cadeias de Markov regulares.

Em síntese, cadeias de Markov, então, são processos que envolvem um determinado sistema e suas transições de estados, sendo possível estabelecer uma probabilidade de o sistema encontrar-se em um estado específico no futuro a partir apenas de seu estado atual. Essa classe de cadeias de Markov são conhecidas, também, como cadeias de Markov sem memória, uma alusão ao fato de o passado ser irrelevante.

Porém, há , também, diversas outras classes de cadeias de Markov sendo uma a ser destacada, as cadeias de Markov de tamanho variável ou cadeias de Markov de ordem variáveis ou, ainda, cadeias de Markov baseadas em

contexto.

Cadeias de Markov baseadas em contexto

As cadeias de Markov baseadas em contexto (ou de ordem variável), diferem das classes de Markov estudadas na seção anterior no que tange ao passado: enquanto estas são classificadas como cadeias de Markov sem memória, aquelas, por sua vez, são conhecidas como cadeias de Markov com memória. Em outras palavras, cadeias de Markov baseadas em contexto estabelecem a probabilidade de um estado futuro com base em seu histórico de estados, e não apenas em seu estado atual.

Contexto, portanto, é a sequência de estados que já ocorreu. Aplicando isso à ideia de compressão de textos, o contexto de um caractere que ainda não foi codificado (compactado) é toda a sequência de caracteres que já foram processados. Dessa forma, tenta-se estabelecer a probabilidade de um caractere ocorrer no texto dado que outros já ocorreram. Por exemplo, em textos em que não haja uma aleatoriedade de caracteres e palavras, como são os textos escritos em linguagem natural tal qual o Português, para uma certa sequência de caracteres “paralelep”, é mais provável que o próximo caractere da sequência seja o “í” aos caracteres “d”, “t”, “x”, entre outros. Neste exemplo, o estado atual da cadeia seria a sequência “paralelep” e o estado futuro seria o estado atual concatenado ao próximo caractere mais

provável “í”. Porém, se fosse levado em conta apenas a cadeia de Markov sem memória, teríamos que prever a probabilidade de um estado futuro dado apenas o estado atual. Neste mesmo exemplo, o atual estado seria apenas o caractere “p” – último da sequência - e a precisão em dizer que o estado futuro seria “pí” seria menor, pois o contexto agrega informação.

Outro ponto a destacar sobre o contexto é o seu tamanho. Um modelo baseado em contexto pode trabalhar com seu tamanho fixo, onde possui um número fixo de caracteres, ou variável. Segundo Hirschberg (1992, p. 3), um modelo que utiliza um contexto de tamanho n (n caracteres prévios) é dito modelo de contexto de ordem n, com n sendo número inteiro. Se $n = 0$, nenhum contexto é usado e a compressão do caractere é feita um caractere por vez. Se $n = 1$, um caractere lido é usado para codificar o próximo. Se $n = 2$, dois caracteres lidos codificam o próximo e assim sucessivamente.

É possível implementar um modelo misto que agrupa quantos modelos de ordem n forem necessários. Dessa forma, um caractere a ser codificado é submetido primeiramente a um modelo de contexto de ordem 3, por exemplo. Se esse modelo não for capaz de estabelecer uma probabilidade, então é submetido a um modelo de ordem 2. Se o mesmo ocorrer, ele pode ser submetido, sucessivamente, até o modelo de ordem -1. Quando a ordem for igual a -1 então o caractere não foi codificado nenhuma vez e, portanto, será estabelecida a sua primeira ocorrência com probabilidade 1. Assim, um modelo misto sempre inicia a codificação baseada nos contextos de ordem

mais elevada e diminui até que seja possível codificar em ordens mais baixas. Ou seja, se para um dado caractere $\boldsymbol{S}$ a ser compactado não for possível encontrar uma ocorrência já computada de $\boldsymbol{n}$ caracteres seguidos de $\boldsymbol{S}$, isto é, um contexto $\boldsymbol{C}$ de ordem $\boldsymbol{n}$ para o caractere $\boldsymbol{S}$, então o modelo de contexto variável verificará se existe alguma ocorrência no contexto dos $\boldsymbol{n-1}$ caracteres (mais à direita) seguidos do caractere $\boldsymbol{S}$. Esse processo decremental do contexto sempre será realizado quando um contexto de ordem $\boldsymbol{n}$ não for capaz de estabelecer uma probabilidade para o caractere $\boldsymbol{S}$**.** O limite dessa operação decremental ocorre com $\boldsymbol{n = -1}$, quando o caractere $\boldsymbol{S}$ ocorrer a primeira vez.

Além disso, os dados sobre os contextos, que serão utilizados para estabelecer as probabilidades dos caracteres do arquivo a ser codificado, podem ser estáticos ou dinâmicos. Serão estáticos quando o compressor possuir, de antemão, uma tabela de contextos que pode ser utilizada em qualquer arquivo a ser compactado. De forma óbvia, neste caso, o descompressor também deverá possuir a tabela de contextos fixa. Porém, esses dados sobre os contextos podem ser construídos de forma dinâmica (adaptativa) e particular para cada arquivo que for compactado durante a própria compactação. A vantagem do modelo estático sobre o modelo adaptativo se dá por uma maior velocidade de compactação, enquanto a vantagem do modelo adaptativo é ser mais eficaz na taxa de compressão.

Algoritmos que aplicam o modelo de contexto

Cada problema pode ser resolvido com variações do modelo baseado em contextos. Algumas abordagens usam como contexto a ideia de palavra, que é uma sequência de caracteres alfabéticos. As cadeias de Markov de contexto também são muito utilizadas na biologia, para realizar mapeamento genético e estabelecer probabilidades de encontrar uma região de origem para determinado gene, entre diversas outras aplicações possíveis.

Na compressão de textos, as cadeias de Markov baseadas em contexto são utilizadas em um algoritmo chamado *Lempel-Ziv Markov Chain Algorithm (LZMA)*, que será estudado em breve.

LZMA

O LZMA é um algoritmo de compressão e descompressão de dados desenvolvido e mantido pelo russo *Igor Pavlov* no ano de 1999. Para uso geral, foi disponibilizado dentro de um programa de compactação de arquivos denominado *7-zip*, desenvolvido pelo mesmo autor, que engloba outros algoritmos de compressão e que possui o próprio formato de arquivo *7z*, além de ser um programa livre (*free*) e de código aberto (*open source*). O *7-zip* por ser de código aberto está disponível no *link* http://www.7-zip.org/sdk.html para qualquer pessoa que tenha o interesse em estudá-lo, modificá-lo e redistribuí-lo.

De acordo com *Pavlov* (2016), o algoritmo LZMA é um algoritmo de alta

taxa de compressão de dados, altamente eficiente na velocidade de descompressão e recomendado para ser embarcado em dispositivos de *hardware*. Em seu documento de especificação técnica chamado *lzma-specification.txt*, disponibilizado por meio do link http://www.7-zip.org/a/lzma-specification.7z, *Pavlov* descreve em termos gerais como funciona seu algoritmo de compressão LZMA, além de dizer que o LZMA é uma combinação dos algoritmos LZ77 (levemente modificado) e Range Encoding, uma outra técnica de compressão (bem como a codificação Huffman), mas que não foi abordado neste trabalho. Além dessa combinação, o LZMA utiliza, entre a aplicação da codificação LZ77 e a aplicação do Range Encoding, um modelo probabilístico baseado em cadeias de Markov de contexto para que o compressor Range Encoding obtenha melhores taxas de compactação.

A dificuldade, entretanto, está situada na forma como o algoritmo LZMA foi escrito. Depois de sucessivas modificações iniciadas ainda em 1999 e continuamente realizadas até 2017, de acordo com o histórico mantido por Pavlov no link http://www.7-zip.org/history.txt, e devido ao estilo de programação do autor, o algoritmo de compressão LZMA não possui uma estrutura que facilite a leitura e a compreensão de como funciona. Assim como ocorre no BZIP2, é um programa altamente modularizado, com um uso de técnicas de programação que visam a eficiência da execução ao entendimento do programa por um terceiro.

Devido a isso não foi possível (no prazo deste trabalho) estabelecer com precisão como os algoritmos LZ77 e Range Encoding são combinados com as Cadeias de Markov de Contexto a fim de promover um processo de compressão e descompressão tão eficiente quanto é prometido pelo autor Pavlov. Assim, as simulações do LZMA tornam-se inviáveis de serem realizadas uma vez que apenas o algoritmo LZ77 que o compõe já fora estudado, desenvolvido e simulado neste projeto.

Uma observação importante: o algoritmo LZMA desenvolvido por Pavlov é utilizado, com modificações, em outras aplicações como os programas XZ Utils e LZMA Utils, ambos desenvolvidos pelo Projeto Tukaani (The Tukaani Project). Todavia, de acordo com o projeto, o programa LZMA Utils foi deprecado em favor de um novo programa denominado XZ Utils (que é totalmente compatível com o algoritmo LZMA), porém incorporando um novo algoritmo chamado LZMA2.

Referências

GRISTEAD, C. M.; SNELL, J. L. *Introduction to Probability.* Edição do Autor, 2006. Disponível em <https://math.dartmouth.edu/~prob/prob/prob.pdf>. Acessado em 27/06/2017.

HIRSCHBERG, D.; LELEWER, D. A. *Context Modeling for Text Compression.* Edição do Autor, 1992.

ROSS, SHELDON. *Probabilidade: um curso moderno com aplicações.* 8.ed. Porto Alegre: Bookman, 2010.

SALOMON, D. *Data compression – the Complete Reference*. New York: Springer, 2007.

THE TUKAANI PROJECT. *XZ Utils*. Disponível em <https://tukaani.org/xz/>. Acessado em 22/08/2017.

WEBER, R. *Markov Chains.* Edição do Autor, 2011. Disponível em <http://www.statslab.cam.ac.uk/~rrw1/markov/M.pdf>. Acessado em 27/06/2017.

7-Zip. *LZMA SDK (Software Development Kit)*, 2016. Disponível em <http://www.7-zip.org/sdk.html>. Acessado em 11/07/2017.