Estudo Dirigido:

Entropia da Informação e Compressão de Dados

Caio Guedes de Azevedo Mota - 2018054990

Gabriel Victor Carvalho Rocha - 2018054907

Guilherme Bezerra dos Santos - 2018055040

Lucas Mariani Paiva Caldeira Brant - 2018054826

1. Entropia da Informação e Compressão de Dados

Entropia, no quesito de teoria da informação, mede o grau médio de incerteza de fontes de informação, ou a possibilidade de obter informação a partir da ocorrência de um processo. A fórmula de entropia da informação, segundo o artigo de C. E. Shannon, *A Mathematical Theory of Communication*, é a seguinte:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

Dado um número de estados possíveis, com probabilidades iguais entre esses estados, a entropia resultante é o logaritmo da quantidade de estados (pode ser obtido pela fórmula a partir de simplificação matemática com $p_i = \frac{1}{n}$).

O *Teorema da Codificação da Fonte* de Shannon é um teorema acerca da entropia H(X) de uma fonte X. Ele afirma, informalmente, que N variáveis aleatórias de uma fonte X, cada uma com entropia H(X), podem ser comprimidas em aproximadamente $N * H(X)$bits. Caso as variáveis sejam comprimidas em menos bits que isso, é quase certo que informação será perdida. No entanto, ao comprimir em mais de $N * H(X)$bits, o risco de perda de informação é insignificante.

Um exemplo é o problema das 12 esferas, abordado no estudo dirigido anterior. Considerando que cada pesagem possui 3 resultados possíveis, a informação de uma pesagem pode ser codificada em um bit ternário (ou trit). Considerando os 24 estados possíveis do problema: $H(X) = log_3 24$, e $N = 1$(um evento do problema).

Portanto, $N * H(X) = log_3 24 \approx 2.89278$. O menor inteiro maior que esse número é 3, portanto 3 trits são necessários para comprimir a informação do problema das 12 esferas.

Ao mudar a base, $N * H(X) = log_2 24 \approx 4.58496$, cujo menor inteiro maior é 5 (ou seja, 5 bits binários necessários sem perda de informação).

2. Estimativa da Entropia por Compressão de Arquivos

A entropia de um sistema pode ser estimada usando compressão de arquivos. Ao salvar a configuração de um sistema em um arquivo e comprimir esse arquivo adequadamente, pode-se estimar a entropia do sistema calculando a razão entre o tamanho do arquivo comprimido e o tamanho do arquivo original. Um simples experimento seria armazenar a configuração de um sistema num arquivo, comprimir ele com um método de compressão de preferência (por exemplo o GZIP, que usa uma mistura de compressão LZ77 com codificação de Huffman), e fazer a razão dos tamanhos.

Algoritmos de compressão resultam em um limite superior para a entropia, logo, a entropia calculada através de algoritmos de compressão será uma superestimação da entropia real. Dito isso, o algoritmo usado no artigo (que é baseado em algoritmos LZ), convergiu para poucas porcentagens de distância da entropia esperada.

Conceitualmente, a redundância de informação armazenada em uma simulação é altamente correlacionada com a entropia do sistema físico sendo simulado. Sendo assim, algoritmos de compressão com maior otimização na redução de redundância são mais adequados para estimar o valor da entropia.

É importante notar que esse método, embora adequado para algumas comparações, não é viável ou adequado para estimar toda e qualquer entropia. Pode não funcionar em casos que a compressão original é feita com perda de dados, ou com métodos de compressão que não funcionam bem com correlações de longo alcance.