Data Compression

Enrico Marchionni
enrico.marchionni@studio.unibo.it

December 10, 2024

Abstract

Data compression is intended as the practice of reducing the size of binary digital data. It could be considered as a procedure that takes a bit-stream in input and returns another bit-stream as output. The output stream may be of equal length or shorter than the input.

The key to understand data compression is to discuss the distinction between data and information. It can be said that data is how information is represented[1]. In simple terms, data can be compressed because its original representation is not the shortest possible. The goal of data compression is to reduce data by maintaining the same information.

The counterpart is that in our time data is intrinsically redundant. And this redundancy is needed. So data compression isn't only a procedure that goes from a bit-stream to another one not longer, but it requires also another procedure that regenerates the original bit-stream of data, necessary for practical use, from the previously given output bit-stream of information.

…

[1]ex. the number 0 can be expressed in binary as a sequence of a certain number of zeros, from 1 to ∞, and we know that calculators use at least 8 bits, let's say n (considering it as a multiple of 8), to represent an integer number. So at the end $n-1$ bits are redundant in the 0 representation on a calculator.

Contents

Chapter 1

Information Theory

In 1948, Shannon[1], while working at the Bell Telephone Laboratories, published "A Mathematical Theory of Communication" [Sha48], a seminal paper that marked the birth of information theory. In that paper, Shannon defined the concept of "information" and proposed a precise way to quantify it-in his theory, the fundamental unit of information is the bit.

Moreover, this discipline plays behind the concepts of entropy, randomness and data compression, all topics that will be discussed later on.

1.1 Quantifying Information

For what concerns data compression, information of theory has developed a usable measure of the information we get from observing the occurrence of an event having probability p. Therefore information is defined in terms of the probability.

The information measure $I(p)$ has to match the following axioms (from [Car07]):

- Information is non negative: $I(p) \geq 0$.
- If an event has probability 1, we get no information: $I(1) = 0$.
- If two independent events occur (whose probability is the product of their individual probabilities), then the information we get from observing the events is the sum of the two computed individually: $I(p_1 \cdot p_2) = I(p_1) + I(p_2)$.
- Information measure must be continuous and monotonic (slight changes in probability should result in slight changes in information).

Considering the previous properties as axioms it can be said that: $I(p^2) = I(p \cdot p) = I(p) + I(p) = 2 \cdot I(p)$. Thus: $I(p^n) = n \cdot I(p)$ (by induction). Then: $I(p) = I(p^{(\frac{1}{m})^m}) = m \cdot I(p^{\frac{1}{m}})$, so $I(p^{\frac{1}{m}}) = \frac{1}{m} \cdot I(P)$, therefore: $I(p^{\frac{n}{m}}) = \frac{n}{m} \cdot I(p)$. In general, considering r as a real number: $I(p^a) = a \cdot I(p)$.

From this analysis it was discovered that:

$$I(p) = -\log_b p \; (= \log_b \frac{1}{p}) \tag{1.1}$$

[1] Claude Elwood Shannon (1916–2001) was an American mathematician, electrical engineer, computer scientist, cryptographer and inventor known as the "father of information theory".

Where: $p = b_1^{\log_{b_1} p}$ and therefore: $\log_{b_2} p = \log_{b_2} b_1^{\log_{b_1} p} = \log_{b_2} b_1 \cdot \log_{b_1} p$. So: $\log_{b_2} b_1$ is a constant, a scaling factor. From another point of view it is a simple change in the unit of measurement.

For this reason:

$$I(p) = -\log_2 p \tag{1.2}$$

Equation 1.2 is the same expression of Equation 1.1 where the unit of measurement is called bits (look at Table 1.1). Equation 1.1 was first introduced by Hartley[2] in 1928 trying to measure uncertainty, without talking about probability, and lately reviewed by Shannon.

Unit of measurement	Base
bit (or shannon)	2
trit	3
nat (natural unit of information)	e
hartley (or dit)	10

Table 1.1: Information units of measurement

Example 1. *Let's talk about flipping a fair coin n times. It gives us:* $-\log_2 \frac{1}{2}^n = \log_2 2^n = n \cdot \log_2 2 = n$ *bits of information. In fact a sequence of heads (coded as* 1*) and tails (coded as* 0*) could be expressed as:* $010010111\ldots$*, these are the* n *bits of information.*

[2]Ralph Vinton Lyon Hartley (1888-1970) was an American electronics researcher. He invented the Hartley oscillator and the Hartley transform, and contributed to the foundations of information theory.

Chapter 2

Entropy

Entropy is a concept that was explained in many fields. Previously defined by Clausius[1] and Boltzmann[2] was later used by Shannon. It is believed that these three definitions are indeed equivalent although no formal proof of this is available (as discussed in [Ben19]).

2.1 Quantifying Entropy

Here is how Shannon introduced the measure of Information:

> Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \ldots, p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or how uncertain we are of the outcome?

If there is such a measure, say, $H(p_1, p_2, \ldots, p_n)$[3], it is reasonable to require of it the following properties:

- H should be continuous in the p_i.
- If all the p_i are equal, $p_i = \frac{1}{n}$ then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
- If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

Then Shannon proved that the only H satisfying the three assumptions above has the form:

$$H = -K \sum_{i=1}^{n} p_i \ln p_i \tag{2.1}$$

Equation 2.1 includes a constant K, in the Shannon article it is any constant. In application to thermodynamics K turns into Boltzmann Constant. It is simply a scaling factor. Note that

[1] Rudolf Julius Emanuel Clausius (1822–1888) was a German physicist and mathematician and is considered one of the central founding fathers of the science of thermodynamics.

[2] Ludwig Eduard Boltzmann (1844–1906) was an Austrian physicist and philosopher. His greatest achievements were the development of statistical mechanics and the statistical explanation of the second law of thermodynamics.

[3] Where H refers to Hartley.

if K is $\frac{1}{\ln b}$ or equivalently $\log_b e$, the formula, considering only K and the logarithm, becomes $\log_b e \cdot \ln p$ that is the same of $\log_b e^{\ln p}$ that can be simply written as $\log_b p$.

So H can be simply reformulated as:

$$H(P) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.2}$$

Where $P = p_1, p_2, \ldots, p_n$ is the distribution of probability considered. remind that in Equation 2.2 base 2 could be a general base b and it can be simply view as a simple change in the unit of measurement (as it was seen in Table 1.1).

An intuitive way to explain the origin of this formula is now discussed. We want to obtain the average amount of information from each symbol we see in a stream. Let's suppose we start from n symbols $a_1, a_1, \ldots, a_n$. A stream of these symbols is provided with probabilities $p_1, p_1, \ldots, p_n$ respectively. As it was seen in Equation 1.2 for a symbol a_i we get $-\log_2 p_i$ information. In a long run, say N observations, we will see (approximately) $N \cdot p_i$ occurrences of the symbol a_i. Thus in the N independent observations, we will get total information of:

$$I = -\sum_{i=1}^{n} (N \cdot p_i) \log_2 p_i \tag{2.3}$$

So then, from Equation 2.3 the average information is:

$$\frac{I}{N} = -\frac{1}{N} \sum_{i=1}^{n} (N \cdot p_i) \log_2 p_i = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.4}$$

At this point we get Equation 2.4 that is the same as Equation 2.2. Furthermore, it is shown in Equation 4.3 that $H(P)$ is bounded (for further information see [Car07]):

$$0 \leq H(P) \leq \log_2 n \tag{2.5}$$

Example 2. *Returning to the example of the coin, in Figure 2.1 it is shown an example of the entropy in function of the probability of heads or tails when flipping a fair coin.*

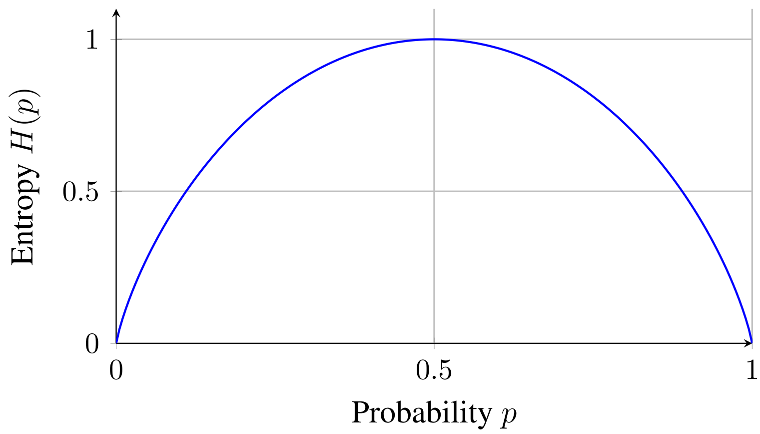

Figure 2.1: Graph of entropy $H(p) = -p\log_2(p) - (1-p)\log_2(1-p)$ for a fair coin toss.

Chapter 3

Randomness

Compression, logically, can be interpreted as the removal of redundancy. The compressed data therefore has no structure and cannot be distinguished from random data; in fact, it is random ([Sal07]).

3.1 Kolmogorov

Definition 3.1.1. ***Kolmogorov complexity*** *of a binary sequence is the length of the shortest binary program that generates that sequence on a universal Turing machine ([Kol68]).*

The concept essentially asserts that a binary sequence is considered random if there is no algorithm shorter than the sequence itself that can generate it.

That is related to the concept of incompressibility in algorithmic randomness: a random sequence cannot be compressed into a shorter representation than its original size.

So, loosely speaking, the randomness (or Kolmogorov complexity) of a finite sequence is equal to its shortest description.

It is known that the Kolmogorov complexity is not computable.

3.2 Martin-Lof

Martin-Lof in Algorithmic Randomness and Complexity ([Mar66]) shows that the random elements as defined by Kolmogorov possess all conceivable statistical properties of randomness.

He also extended the definition for random elements with three approaches to the definition of algorithmic randomness for infinite sequences:

Definition 3.2.1. *The* ***computational paradigm****: Random sequences are those whose initial segments are all hard to describe, or, equivalently, hard to compress.*

Definition 3.2.2. *The* ***measure-theoretic paradigm****: Random sequences are those with no "effectively rare" properties. If the class of sequences satisfying a given property is an effectively null set, then a random sequence should not have this property. This approach is the same as the stochastic paradigm: a random sequence should pass all effective statistical tests.*

Definition 3.2.3. *The* ***unpredictability paradigm****: This approach stems from what is probably the most intuitive conception of randomness, namely that one should not be able to predict the next bit of a random sequence, even if one knows all preceding bits, in the same way that a coin toss is unpredictable even given the results of previous coin tosses.*

Taken from [DH10].

The previous citations aim to observe that the idea of Kolmogorov that random generators didn't exist was lately reviewed and while remaining true, extended to infinite sequences.

3.3 Random Sequences

A random sequence should satisfy three conditions:

- The sequence follows a uniform distribution.
- Each element of the sequence is independent of each other.
- The rest of the sequence can not be predicted from any sequence.

Random numbers can be divided into two categories: true random numbers and pseudo-random numbers. True random number generators (RNGs) are composed of two parts: entropy source and algorithm post-processing. Pseudo random number generators (PRNGs) take a seed as input and generate an output sequence by function.

Chapter 4

Data Compression

4.1 Techniques

The compression techniques can be divided in two main types "lossless" and "lossy". The first category is the most similar to the theory and the most practical intuition of it. Lossless compression ratios are generally in the range of 2:1 to 8:1. Lossy compression, in contrast, works on the assumption that the data doesn't have to be stored perfectly. Most information can be simply thrown away; the data will still be of acceptable quality. This technique is frequently used in image and video compression, where a group of pixels can be approssimated into a single value. In this case the ratio can be of orders greater. In conclusion losseless compression has a lower ratio it preserves all information that can be reload back to the original data, on the other hand, lossy compression has a bigger ratio but it discards some information and doesn't generate the same data when it is reloaded.

4.2 Comparison of Algorithms

They are useful when done on algorithms of the same compression technique: lossless or lossy. Measurements have to be considered as averages in practice, because they are different for every single sample so their value is an approximation.

4.2.1 Compression Ratio

An important factor when considering data compression is the "compression ratio". It is simply the sample size reduction factor.

$$compressionratio = \frac{uncompressedfilesize}{compressedfilesize} \tag{4.1}$$

4.2.2 Compression Speed

It is important is specific use cases, usually when the network is involved.

$$compressionspeed(MB/s) = \frac{uncompressedfilesize(MB)}{compressiontime(s)} \tag{4.2}$$

4.2.3 Decompression Speed

As the compression speed, it is important is specific use cases, usually when the network is involved.

$$decompressionspeed(MB/s) = \frac{uncompressedfilesize(MB)}{uncompressiontime(s)} \tag{4.3}$$

Chapter 5

Compression Tools

…

Bibliography

[Ben19] Arieh Ben-Naim. "Entropy and Information Theory: Uses and Misuses". In: *Entropy* 21.12 (2019). ISSN: 1099-4300. DOI: 10.3390/e21121170. URL: https://www.mdpi.com/1099-4300/21/12/1170.

[DH10] Rodney G. Downey and Denis R. Hirschfeldt. *Algorithmic Randomness and Complexity*. Springer, 2010. URL: https://books.google.it/books?id=FwIKhn4RYzYC&lpg=PR3&ots=_nbPVe6IUi&dq=randomness%20in%20computer%20science&lr&pg=PA226#v=onepage&q&f=false.

[Car07] Tom Carter. "An introduction to information theory and entropy". In: *Complex systems summer school, Santa Fe* (2007).

[Sal07] D. Salomon. *A Concise Introduction to Data Compression*. Undergraduate Topics in Computer Science. Springer London, 2007. ISBN: 9781848000728. URL: https://books.google.it/books?id=mnpeizY0btYC.

[Kol68] A. N. Kolmogorov. "Three approaches to the quantitative definition of information". In: *International Journal of Computer Mathematics* 2.1-4 (1968), pp. 157–168. DOI: 10.1080/00207166808803030. eprint: https://doi.org/10.1080/00207166808803030. URL: https://doi.org/10.1080/00207166808803030.

[Mar66] Per Martin-Löf. "The definition of random sequences". In: *Information and Control* 9.6 (1966), pp. 602–619. ISSN: 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(66)80018-9. URL: https://www.sciencedirect.com/science/article/pii/S0019995866800189.

[Sha48] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.