Atypical Information Theory for BIG DATA

Elyas Sabeti, Anders Høst-Madsen and Chad Walton

Department of Electrical Engineering and Department of Medicine, University of Hawaii

Objectives

Our main Goals are:

- Defining Atypicality.
- Finding Atypical subsequences.
- Finding the probability that a sequence of length l is classified as atypical.
- Finding the probability that a given sample is a part of an atypical subsequence of any length.

Introduction

Information theory is generally a theory of typicality. For example, compressing data using the Asymptotic Equipartion Property (AEP) can be done by throwing away all sequences that are not typical. Our perspective in this paper is that the value of data lies not in these typical sequence, but in the atypical sequences. With the enormous amount of data generated with modern technology and available through data networks and the internet ("Big Data"), our perspective is that most of this data is background noise. What is valuable are the outliers from this background noise. What is an atypical sequence? Consider throwing a fair coin. If we get a sequence a 100 consecutive heads, we would be surprised. If we were in a casino, security would scrutinize our gambling. Yet, a sequence of 100 consecutive heads does not contradict the laws of probability for a fair coin. In fact, for a fair coin the sequence of 100 consecutive heads has exactly the same probability as any other sequence. Now suppose we instead used a biased coin with probability 0.99 of getting a head. A sequence of 100 heads would not be unexpected; in fact a sequence with 99 heads would be the most likely. The causes of the two outcomes are different. Yet, for a casino both sequences would be worthy of scrutiny. We call the first type of sequence intrinsically atypical, while the second type will be called extrinsically atypical.

Definition

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

Finding Atypical Subsequces in Binary IID Sequences

Consider a sequence of random variables $\{x[n], n = -\infty, \dots, \infty\}$ from a finite alphabet $\mathcal{A}$. The sequence is generated according to a probability law $\mathcal{P}$ (at first vaguely specified). In this sequence is embedded (infrequent) finite subsequences $\mathcal{X}_i = \{x[n], n = n_i, \dots, n_i + l_i - 1\}$ from the finite alphabet $\mathcal{A}$. which are generated by an alternative probability law $\tilde{\mathcal{P}}_\theta$ which is unknown.

In the following we consider a very simplified model where the typical sequences are iid (identically, independent distributed) so that the probability law $\mathcal{P}$ is specified by the single parameter $p = P\{x[n] = 1\}$. The alternative law $\tilde{\mathcal{P}}_\theta$ is also iid with $\theta = P\{x[n] = 1\} \neq p$. According to universal source coding, in a sequence with k ones, the total code length of the universal code for atypical sequences is approximately

$$L_{\hat{p}}(l) = lH(\hat{p}) + \frac{3}{2}\log(l) + \tau \tag{1}$$

Where

$$\hat{p} = \frac{1}{l}\Sigma X_i \tag{2}$$

The code length for the sequence coded according to the optimum code for the the typical probability law $\mathcal{P}$ is approximately

$$l(\hat{p}\log(\frac{1}{p}) + (1-\hat{p})\log(\frac{1}{1-p})) \tag{3}$$

So difference between these to code length will leads to the following hypothesis test for atypical sequences

$$D(\hat{p}||p) > \frac{\tau + \frac{3}{2}\log(l)}{l} \tag{4}$$

Theorems

In this section, we will answer to two of our principal questions: What is the probability of a sequence of length l be classified as atypical and what is the probability that a given sample be a part of an atypical subsequence of any length.

Theorem 1. Consider an iid $\{0,1\}-$sequence with $P(X=1) = p$. The probability P_A that a sequence of length l is classified as atypical according to (4) is bounded by

$$P_A \leq 2^{-\tau+1}\frac{1}{l^{3/2}}K(l,\tau),\ \forall\tau : \lim_{l\to\infty} K(l,\tau) = 1 \tag{5}$$

Theorem 2. Consider the case $p = \frac{1}{2}$. The probability $P_A(X_n)$ that a given sample X_n is part of an atypical subsequence of any length is upper bounded by

$$P_A(X_n) \leq (K_1\sqrt{\tau} + K_2)2^{-\tau} \tag{6}$$

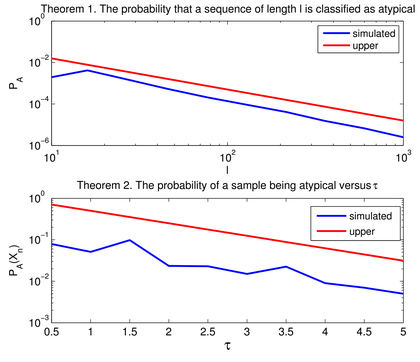

Figure 1: The top figure shows the probability of a sequence of length l being atypical with $\tau = 1, p = \frac{1}{2}$. The bottom figure shows the probability of a given sample being part of an intrinsically atypical sequence of any length, with $p = \frac{1}{2}$.

Universal Source Coding

A natural way to extend the simple example to non-IID sequences is universal source coding. Here we will use the Context Tree Weighing (CTW) algorithm. Consequently, for every bit of the data we calculate

$$\Delta L(n) = \min_l L_A(\mathcal{X}(l)) - L_T(\mathcal{X}(l)) \tag{7}$$

Where $L_A(\mathcal{X}(l))$ and $L_T(\mathcal{X}(l))$ are the length of atypical sequence and typical sequence, respectively.

Application Example

As application we take HRV (heart rate variability). Here is the result of the algorithm on HRV:

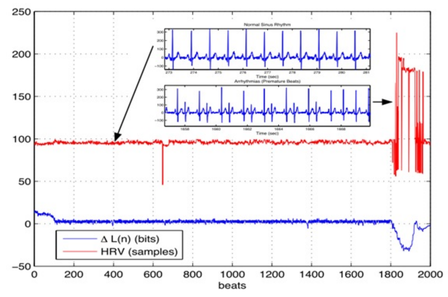

Figure 2: THRV signal with premature beats.

References

[1] J. Rissanen, "A universal prior for integers and estimation by minimum description length",The Annals of Statistics, no. 2, pp. 416Ð431, 1983.

Contact Information

- Email: sabeti@hawaii.edu
- Phone: +1 (808) 799 8088

College of Engineering, University of Hawai'i at Mānoa — University of Hawai'i Mānoa