REVISITED BOX COUNTING TECHNIQUE IN BAYESIAN SENSE

Václav Hubata–Vacek [1], Jaromír Kukal [1]

[1]CTU in Prague, Faculty of Nuclear Sciences and Physical Engineering
Department of Software Engineering in Economics
Břehová 7, 115 19 Prague 1
Czech Republic
hubatvac@fjfi.cvut.cz

Abstract:

Keywords: *EEG, Alzheimer's disease, permutation entropy, unbiased estimation, hash table*

1 Introduction

2 Multinomic Distribution

Multinomic distrubution model plays main role in investigation of point set structures. Let $n \in \mathbb{N}$ be number of distinguish events. Let $p_j > 0$ be probability of j^{th} event for $j = 1, ..., n$ satisfying $\sum_{j=1}^{n} p_j = 1$. Then random variable j has multinomic distribution $Mul(p_1, ..., p_n$. After realization of multinomic distribution sample of size $N \in \mathbb{N}$, we can count the events and obtain $N_j \in \mathbb{N}_0$ as number of j^{th} event occurences for $j = 1, ..., n$ satisfying $\sum_{j=1}^{n} N_j = N$. Therefore, we define number of various events in sample as $K = \sum_{N_j > 0} 1 \leq min(n, N)$. Remembering Hartley and Shannon entropies definitions

$$H_0 = \ln n, \tag{1}$$

$$H_1 = -\sum_{j=1}^{n} p_j \ln p_j, \tag{2}$$

we can perform naive or rather naive estimation of them as

$$H_{0,\mathrm{NAIVE}} = \ln K, \tag{3}$$

$$H_{1,\mathrm{NAIVE}} = -\sum_{j=1}^{n} \frac{N_j}{N} \ln \frac{N_j}{N}. \tag{4}$$

The main disadvantage of naive estimation is their biases. Random variable $K = \{1, ..., n\}$ is upper construined by n, then $EH_{0,\mathrm{NAIVE}} = E \ln K < E lnn = \ln n = H_0$. Therefore, naive estimate of Hartley entropy $H_{0,\mathrm{NAIVE}}$ is negative biased. On the other hand, traditional Box Counting Technique is based on this estimate because we plot logarithm of covering element number $C(a) \in \mathbb{N}$ against logarithm of covering element size $a > 0$ and then estimate their dependency in linear form $\ln C(a) = A_0 - D_{0,\mathrm{NAIVE}} \ln a$. Recognizing equivalence $C(a) = K$, we obtain $\ln C(a) = \ln K = H_{0,\mathrm{NAIVE}}$ and then $H_{0,\mathrm{NAIVE}} = A_0 - D_{0,\mathrm{NAIVE}} \ln a$. Defining $D_{0,\mathrm{NAIVE}}$ as estimate of capacity dimension and recognizing the occurence of $H_{0,\mathrm{NAIVE}}$ in Box Counting procedure, we are not suprised to be victims of the bias of Hartley entropy estimate.

3 Bias of Shannon entropy

Similar situation is the case of Shannon entropy estimation. There are several approaches how to declare the bias of $H_{1,\mathrm{NAIVE}}$ to be closer to Shannon entropy H_1.

TODO (Miller, Harris, modifacions)

Finally, we can estimate information dimension according to relation

$$H_{1,\mathrm{EST}} = A_1 - D_{1,\mathrm{EST}} \ln a \tag{5}$$

where $H_{1,\mathrm{EST}}$ is any estimate of H_1. Therefore, we can also estimate Hansdorf dimension D_{H} using inequalities $D_1 \leq D_{\mathrm{H}} \leq D_0$ and then also supposing $D_{1,\mathrm{EST}} \leq D_{\mathrm{H}} \leq D_{0,\mathrm{EST}}$ for any "good" estimates $D_{0,\mathrm{EST}}$, $D_{1,\mathrm{EST}}$ of capacity and information dimensions. Next section is oriented to Bayesian estimation of H_0, H_1 for $D_{0,\mathrm{EST}}$ and $D_{1,\mathrm{EST}}$ evaluations.

4 Bayesian estimation of Hartley entropy

Having again $n, N \in \mathbb{N}$ as number of possible events and sample size we can suppose uniform distribution of random vector $\vec{p} = (p_1, ..., p_n)$ satisfying $p_j \geq 0$, $\sum_{j=1}^{n} p_j = 1$. Using properties of multinomic and Dirichlet distributions, we can calculate density $p(K|n,N)$ of random variable $K \in \mathbb{N}$ for $K \leq \min(n,N)$ as

$$p(K|n,N) = prob(\sum_{N_j>0} 1 = K|n, \sum_{j=1}^{n} N_j = N) = \frac{\binom{n}{K}\binom{N-1}{K-1}}{\binom{N+n-1}{n-1}}. \tag{6}$$

When $N \geq K+2$, we can calculate

$$S_{K,N} = \sum_{n=K}^{\inf} p(K|n,N). \tag{7}$$

Using inequality

$$\begin{aligned} p(K|n,N) =& \frac{N!(N-1)!}{K!(K-1)!(N-K)!} \frac{n!(n-1)!}{(n-K)!(n+N-1)!} = \\ =& q(K,N) \frac{n(n-1)...(n-k+1)}{(n+N-1)(n+N-2)...n} \leq q(K,N) \frac{n^K}{n^N} \end{aligned} \tag{8}$$

we can overestimate

$$S_{K,N} \leq \sum_{n=K}^{\inf} q(K,N) n^{K-N} = q(K,N) \sum_{n=K}^{\inf} n^{K-N} < +\inf \tag{9}$$

and then recognize the convergence of infinite series TODO. Having a knowledge of K, N where $N \geq K+2$, we can calculate bayesian density

$$p(n|K,N) = \frac{p(K|n,N)}{S_{K,N}} \tag{10}$$

for $n \geq K$. Therefore, Bayesian estimate of Hartley entropy is

$$H_{0,\text{BAYES}} = EH_0 = \sum_{n=K} \inf p(n|K,N) \ln n > \ln k \tag{11}$$

which is also convergent sum. Substituing $n = K + j$ we obtain equivalent formula

$$H_{0,\text{BAYES}} = \frac{\sum_{j=0} \inf b_j \ln K + j}{\sum_{j=0} \inf b_j} \tag{12}$$

where $b_j = \frac{\binom{K+j}{j}\binom{K+j-1}{j}}{\binom{K+j+N-1}{j}}$
Then TODO
Asymptotic properties of Bayesian estimate for $N \to +\inf$ can be investigated via limits

$$\lim_{N \to +\inf} H_{0,\text{BAYES}} = \ln K, \tag{13}$$

$$\lim_{N \to +\inf} (H_{0,\text{BAYES}} - \ln K) N = K(K+1) \ln 1 + 1/K, \tag{14}$$

$$\lim_{N \to +\inf} (H_{0,\text{BAYES}} - \ln K - \frac{K(K+1)\ln(1+1/K)}{N}) N^2 = TODO, \tag{15}$$

TODO
Therefore

$$H_{0,\text{BAYES}} \approx \ln K + \frac{K(K+1)\ln(1+1/K)}{N} + TODO \tag{16}$$

When K is also large, we can roughly approximate Hartley entropy as

$$H_{0,\text{BAYES}} \approx \ln K + \frac{K+1}{N} \tag{17}$$

which is very similar to Miller correction TODO in the case of Shannon entropy estimation.

5 Bayesian estimation of Shannon entropy

In the case when the number of events n is known, we can perform Bayesian estimation of Shannon entropy as

$$H_{1,\mathrm{n}} = EH_1(K = m) = \sum_{j=1}^{m} \left(\frac{N_j + 1}{N + m} \sum TODO\right) \tag{18}$$

as derived in [TODO]. But when the number of events n is unkonown, we can use k as lower estimate of n and perform final Bayesian estimation as

$$H_{1,\mathrm{BAYES}} = \sum_{n=K}^{\mathrm{inf}} p(n|K, N) H_{1,\mathrm{n}} \tag{19}$$

which is also convergent sum for $N \geq K + 2$. Asymptotic properties of (TODO) for $N \to +\mathrm{inf}$ can be investigated by the same technique as in previous section. Resulting asymptotic formula is

$$H_{1,\mathrm{BAYES}} \approx c_0 + \frac{c_1}{N} + \frac{c_2}{N^2} \tag{20}$$

where

$$c_0 = \sum_{N_j > 0} \frac{N_j}{N} \ln \frac{N}{N_j} TODOc1, c2 \tag{21}$$

Here c_0 is naive estimate of Shannon entropy, c_1/N corresponds with Miller estimate [TODO] but the term c_2/N^2 differs from Harris estimate [TODO]. The main advantage of formulas TODO,TODO is in absence of theoretical probability knowledge.

6 REvisited Cox Counting

7 Experimental parth

8 Conclusion

Acknowledgement: The paper was created with the support of CTU in Prague, Grant SGS11/165/OHK4/3T/14.

References

[1] todo t.*todo*. todo.

A Main function for permutation

```
function
```

B Hash function

```
function
```