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: *unbiased estimation*

1 Introduction

2 Multinomic Distribution and Naive Entropy Estimates

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 $\text{Mul}(p_1, ..., p_n)$. 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 as

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

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

we can perform direct but naive estimation of them as

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

$$H_{1,\text{NAIVE}} = -\sum_{N_j>0} \frac{N_j}{N} \ln \frac{N_j}{N}. \tag{4}$$

The main disadvantage of naive estimates is their biasness. Random variable $K = \{1, ..., n\}$ is upper construined by n, then $\text{E}H_{0,\text{NAIVE}} = \text{E} \ln K < \text{E} \ln n = \ln n = H_0$. Therefore, naive estimate of Hartley entropy $H_{0,\text{NAIVE}}$ 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 $\ln C(a) = A_0 - D_{0,\text{NAIVE}} \ln a$. Recognizing equivalence $C(a) = K$, we obtain $\ln C(a) = \ln K = H_{0,\text{NAIVE}}$ and then $H_{0,\text{NAIVE}} = A_0 - D_{0,\text{NAIVE}} \ln a$. Defining $D_{0,\text{NAIVE}}$ as estimate of capacity dimension and recognizing the occurence of $H_{0,\text{NAIVE}}$ in Box Counting procedure, we are not suprised to be victims of the bias of Hartley entropy estimate.

Similar situation is the case of Shannon entropy estimation. There are several approaches how to declare the bias of $H_{1,\text{NAIVE}}$ to be closer to Shannon entropy H_1. Miller [2] modified naive estimate $H_{1,\text{NAIVE}}$ using first order Taylor expansion, which produces

$$H_{1,\text{M}} = H_{1,\text{NAIVE}} + \frac{K-1}{2N}. \tag{5}$$

Lately, Harris [2] improved the formula to

$$H_{1,\text{H}} = H_{1,\text{NAIVE}} + \frac{K-1}{2N} + \frac{1}{12N^2} \left(1 - \sum_{p_j>0} \frac{1}{p_j}\right) \tag{6}$$

Finally, we can estimate information dimension according to relation

$$H_{1,\mathrm{EST}} = A_1 - D_{1,\mathrm{EST}} \ln a \tag{7}$$

where $H_{1,\mathrm{EST}}$ is any estimate of H_1. Therefore, we can also estimate Hausdorf 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.

3 Bayesian Estimation of Hartley Entropy

We 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

$$\mathrm{p}(K|n,N) = \mathrm{prob}\left(\sum_{N_j>0} 1 = K \middle| n, \sum_{j=1}^{n} N_j = N \right) = \frac{\binom{n}{K}\binom{N-1}{K-1}}{\binom{N+n-1}{n-1}}. \tag{8}$$

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

$$S_{K,N} = \sum_{n=K}^{\infty} p\,(K|n,N). \tag{9}$$

Using inequality

$$\begin{aligned} \mathrm{p}(K|n,N) =& \frac{N!(N-1)!}{K!(K-1)!(N-K)!} \frac{n!(n-1)!}{(n-K)!(n+N-1)!} = \\ =& \mathrm{q}(K,N) \frac{n(n-1)...(n-k+1)}{(n+N-1)(n+N-2)...n} \leq \mathrm{q}(K,N) \frac{n^K}{n^N} \end{aligned} \tag{10}$$

we can overestimate

$$S_{K,N} \leq \sum_{n=K}^{\infty} \mathrm{q}(K,N) n^{K-N} = \mathrm{q}(K,N) \sum_{n=K}^{\infty} n^{K-N} < +\infty \tag{11}$$

and then recognize the convergence of infinite series (11). Having a knowledge of K, N where $N \geq K + 2$, we can calculate bayesian density

$$\mathrm{p}\,(n|K,N) = \frac{\mathrm{p}\,(K|n,N)}{S_{K,N}} \tag{12}$$

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

$$H_{0,\mathrm{BAYES}} = \mathrm{E}H_0 = \sum_{n=K}^{\infty} \mathrm{p}(n|K,N) \ln n > \ln k \tag{13}$$

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

$$H_{0,\mathrm{BAYES}} = \frac{\sum_{j=0}^{\infty} b_j \ln\,(K+j)}{\sum_{j=0}^{\infty} b_j} \tag{14}$$

where $b_j = \frac{\binom{K+j}{j}\binom{K+j-1}{j}}{\binom{K+j+N-1}{j}}$.

Then

$$\begin{aligned} b_0 &= 1, \\ b_1 &= \frac{(K+1)K}{K+N}, \\ b_2 &= \frac{1}{2} \frac{(K+2)(K+1)^2 K}{(K+N+1)(K+N)}, \\ b_3 &= \frac{1}{6} \frac{(K+3)(K+2)^2(K+1)^2 K}{(K+N+2)(K+N+1)(K+N)}, \\ b_4 &= \frac{1}{24} \frac{(K+4)(K+3)^2(K+2)^2(K+1)^2 K}{(K+N+3)(K+N+2)(K+N+1)(K+N)}. \end{aligned} \tag{15}$$

Asymptotic properties of Bayesian estimate for $N \to +\infty$ can be investigated via limits

$$\begin{aligned}
&\lim_{N\to+\infty} H_{0,\mathrm{BAYES}} = \ln K,\\
&\lim_{N\to+\infty} (H_{0,\mathrm{BAYES}} - \ln K)N = K(K+1)\ln(1+1/K),\\
&\lim_{N\to+\infty} \left(H_{0,\mathrm{BAYES}} - \ln K - \frac{K(K+1)\ln(1+1/K)}{N}\right) N^2 =\\
&\qquad \frac{1}{2}\left(K(K+2)(K+1)\left(\ln(K+2) - \ln(K) - 2K\ln(K+1) + K\ln(K+2) + K\ln(K)\right)\right),
\end{aligned} \tag{16}$$

Therefore

$$\begin{aligned}
H_{0,\mathrm{BAYES}} \approx &\ln K + \frac{K(K+1)\ln(1+1/K)}{N} + \\
&\frac{\left(K(K+2)(K+1)\left(\ln(K+2) - \ln(K) - 2K\ln(K+1) + K\ln(K+2) + K\ln(K)\right)\right)}{2N^2}
\end{aligned} \tag{17}$$

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

$$H_{0,\mathrm{BAYES}} \approx \ln K + \frac{K+1}{N} \tag{18}$$

which is very similar to Miller correction [2] in the case of Shannon entropy estimation.

4 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

$$H_{1,\mathrm{n}} = \mathrm{E}H_1(K=m) = -\sum_{j=1}^{m}\left(\frac{N_j+1}{N+m}\left(\psi(N_j+2) - \psi(N+m+1)\right)\right) \tag{19}$$

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

$$H_{1,\mathrm{BAYES}} = \sum_{n=K}^{\infty} \mathrm{p}\left(n|K,N\right) H_{1,\mathrm{n}} \tag{20}$$

which is also convergent sum for $N \geq K+2$. Asymptotic properties of (20) for $N \to +\infty$ can be investigated by the same technique as in previous section. Resulting asymptotic formula is

$$H_{1,\mathrm{BAYES}} \approx c_0 + \frac{c_1}{N} + \frac{c_2}{N^2} \tag{21}$$

where

$$\begin{aligned}
c_0 &= \sum_{N_j>0} \frac{N_j}{N}\ln\frac{N}{N_j},\\
c_1 &= TODO,\\
c_2 &= TODO.
\end{aligned} \tag{22}$$

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.

5 Revisited Box Counting

6 Experimental parth

7 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.

[2] Harris, B., *The statistical estimation of entropy in the non-parametric case*. MRC Technical Summary Report, 1975

Let $\mathbb{Q}_n = \{\vec{q} \in (\mathbb{R}_0^+)^n | \sum_{j=1}^n q_j = 1\}$ be support set for uniform random variable $\vec{p} \in \mathbb{Q}_n$. Then for integer K satisfying $1 \leq K \leq min(n, N)$, the conditional probability is

$$\mathrm{p}(K|n,N) = \mathrm{prob}\left(\sum_{j=1}^{n}(N_j > 0) = K \middle| n, \sum_{j=1}^{n} N_j = N\right). \tag{23}$$

The vector of N_j can be reorganized to begin with positive values. Therefore

$$\mathrm{p}(K|n,N) = \binom{n}{K}\mathrm{prob}\left(N_j > 0 \Leftrightarrow j \leq K \middle| n, \sum_{j=1}^{K} N_j = N\right). \tag{24}$$

Let $\mathbb{D}_{K,N} = \{\vec{x} \in \mathbb{N}^K | \sum_{j=1}^K x_j = N\}$ be domain of $\vec{N} = (N_1, ..., N_K) \in \mathbb{D}_{K,N}$. Using mean value of multinomic distribution over $\mathbb{Q}_n$, we obtain

$$\mathrm{p}(K|n,N) = \binom{n}{K}\mathrm{E}\left(\sum_{\vec{N}\in\mathbb{D}_{K,N}}\binom{N}{N_1,...,N_K}\prod_{j=1}^{K}p_j^{N_j}\prod_{j=k+1}^{n}p_j^0\right) = \binom{n}{K}\mathrm{E}\left(\sum_{\vec{N}\in\mathbb{D}_{K,N}}\binom{N}{N_1,...,N_K}\mathrm{E}\prod_{j=1}^{K}p_j^{N_j}\right). \tag{25}$$

Using generalized Beta function

$$B(\vec{x}) = \int_{\vec{p}\in\mathbb{Q}_m}\prod_{j=1}^{m}p_j^{x_j-1}d\vec{p} = \frac{\prod_{j=1}^m\Gamma(x_j)}{\Gamma(\sum_{j=1}^m x_j)}, \tag{26}$$

we can calculate

$$\mathrm{E}(\prod_{j=1}^{K}p_j^{N_j}) = \frac{\int_{\vec{p}\in\mathbb{Q}_n}\prod_{j=1}^K p_j^{N_j}d\vec{p}}{\int_{\vec{p}\in\mathbb{Q}_n}d\vec{p}} = \frac{\prod_{j=1}^K\Gamma(N_j+1)}{\Gamma(N+n)} \tag{27}$$

Therefore

$$\mathrm{prob}(K|n,N) = \binom{n}{K}\sum_{\vec{N}\in\mathbb{D}_{K,N}}\frac{N!(n-1)!}{\prod_{j=1}^K N_j!}\frac{\prod_{j=1}^K N_j!}{(N+n-1)!} = \binom{n}{K}\sum_{\vec{N}\in\mathbb{D}_{K,N}}\frac{N!(n-1)!}{(N+n-1)!} = \frac{\binom{n}{K}\mathrm{card}(\mathbb{D}_{K,N})}{\binom{N+n-1}{n-1}}. \tag{28}$$

The last question is about the cardinality of $\mathbb{D}_{k,N}$, which corresponds with number of possibilities, how to place N balls into k boxes under assumptions that no box is empty and the balls are identic. We place k balls into k different boxes in the first phase. The rest of $N-k$ balls can be distributed without anz constrains. Therefore

$$\mathrm{card}(\mathbb{D}_{K,n}) = \binom{(N-K)+K-1}{K-1} = \binom{N-1}{K-1}. \tag{29}$$

Resulting formula is (8).

A Hash function

```
function
```