# Empirical Active Information

Eric Holloway

DOI: 10.33014/issn.2640-5652.2.2.holloway.4

Active information is the log ratio of the performance of a search with ($\theta$) and without ($\beta$) information about the target it is searching for, using the notation of self information $I(p) = -\log_2 p$,

$$I_+ = \log_2 \frac{\theta}{\beta} \tag{1}$$

$$= I(\beta) - I(\theta). \tag{2}$$

As proven in Bartlett (2020), active information is a canonical specified complexity, and therefore by Montañez (2018),

$$\Pr(I_+ > \alpha) \leq 2^{-\alpha}. \tag{3}$$

This is posed in an ideal setting where we know both of these performance values.

However, in the empirical setting, although we can perhaps take the baseline as a given value, we must infer the performance of the search with target information by observation. In this case, we will only consider a search across binary strings. All finite domains can be cast as binary strings, so the proof here is general enough for all finite domains, although perhaps not always the most convenient representation.

We've observed $n$ bits and of those $k$ are zeros. We can think of this in terms of a prediction task, where ones correspond to correct predictions and zeros correspond to incorrect predictions.

For sake of demonstration we set

$$k \leq n/2, \tag{4}$$

which we can always guarantee by flipping the prediction if more than half are incorrect.

From the assumption in (4) we define the probabilities

$$p = k/n \tag{5}$$

$$q = 1 - k/n \tag{6}$$

such that $p \leq q$.

We want to figure out what kind of prediction accuracy we can expect on unseen data, and we want to err on the side of caution and not be over optimistic. Stated more formally, what is the Bernoulli distribution probability $\epsilon$ of ones that generated this sequence, based on a certain confidence probability of $\lambda$?

First, here is an upper bound on the binomial quotient tail derived from the binomial cumulative distribution:

$$\sum_{c=0}^{k} \binom{n}{c} p^{pn} q^{qn} \leq \sum_{c=0}^{k} \binom{n}{c} p^c q^{n-c} \leq 1 \tag{7}$$

$$\sum_{c=0}^{k} \binom{n}{c} \leq p^{-pn} q^{-qn} \tag{8}$$

$$= 2^{nH(p)}. \tag{9}$$

The left hand side of (7) is derived by keeping the probability constant during the summation, instead of changing as $c$ is incremented. This guarantees the left hand side is a lower bound on the binomial cumulative distribution, since we've selected the smallest probability in the summation and kept it constant.

The jump from $p^{-pn}q^{-qn}$ in (8) to $2^{nH(p)}$ in (9) is based on the entropy function:

$$H(p) = -p \log_2 p - q \log_2 q \tag{10}$$

$$2^{H(p)} = p^{-p} q^{-q}. \tag{11}$$

Now, let us return to our goal of estimating the accuracy on unseen data. The estimate can be wrong in two ways.

1. Too low. If so, then great!
2. Too high. If so, we want to know the probability the estimate is too high.

So, stated more formally, we want to pick an $\epsilon$ for our estimate such that any Bernoulli distribution with a smaller probability $\beta$ of generating a one, i.e. $0.5 \leq \beta < \epsilon$, will have a probability less than $\lambda$ of generating $k$ or fewer zeros. In other words, $\lambda$ is the probability we've over estimated our model accuracy.

The probability $\lambda$ is thus:

$$\lambda = \sum_{c=0}^{k} \binom{n}{c} (1-\epsilon)^c \epsilon^{n-c} \leq \sum_{c=0}^{k} \binom{n}{c} \epsilon^n \tag{12}$$

$$\leq 2^{nH(p)} \epsilon^n \tag{13}$$

$$\log_2 \lambda \leq n(H(p) + \log_2 \epsilon). \tag{14}$$

The substitution to get the right hand side in (12) is the same sort of reasoning used to get the left hand side in (7), except reversed. In (7) the goal was to get a lower bound

on the cumulative distribution. In (12) the goal is to get an upper bound. So, we pick the largest probability in the summation, $\epsilon^n$, which is larger than $(1-\epsilon)^c\epsilon^{n-c}$ because $\epsilon \geq 0.5$.

The reduction from (12) to (13) is based on the substitution from (9). The use of the logarithm in (14) measures the probability in bits to make the formula a bit easier to manipulate.

Now, we want to derive $\epsilon$ from a predefined value for $\lambda$, for example $\lambda = 2^{-m}$. This means we need to pick a value for $\epsilon$ such that the right hand side of (14) evaluates to $-m$. If we substitute in

$$\epsilon = 2^{-H(p)-m/n}, \tag{15}$$

we achieve just that result,

$$\log_2 \lambda \leq n(H(p) + \log_2(2^{-H(p)-m/n})) \tag{16}$$
$$= n(H(p) - H(p) - m/n) = -m. \tag{17}$$

As long as $nH(p)+m < n$ then from (15) $\epsilon > 0.5$, and we are able to predict better than 50/50 odds. Here we can also see why our very first assumption in (4) is necessary. If $k > n/2$ then since the entropy function $H(p)$ is symmetrical we can substitute in a value for $p$ such that (15) gives us a estimate that is higher than the observed accuracy rate, i.e. $q < \epsilon$, which is a contradiction since the estimate is a lower bound and so should be lower than the observed accuracy.

Now we can finalize our concept of empirical active information per sample by substituting in $\theta = \epsilon$ for (1) and setting $m = 0$,

$$I_+^e = I(\beta) - H(p). \tag{18}$$

And as at the beginning, the probability bound still applies.

$$\Pr(I_+^e > \alpha) \leq 2^{-\alpha}. \tag{19}$$

---

Bartlett, Jonathan (2020). "Active Information is a Specified Complexity Model". In: *Communications of the Blyth Institute* 2.2.

Montañez, George D (2018). "A Unified Model of Complex Specified Information". In: *BIO-Complexity* 2018.