Extensions to the context tree weighting method

Willems, F.M.J.

Published in:

DOI:
10.1109/ISIT.1994.394632

Published: 01/01/1994

Citation for published version (APA):
Abstract — We modify the basic context tree weighting method so that past symbols are not needed, and that the context tree depth is infinite. For stationary ergodic sources we now achieve entropy.

I. Introduction

The context tree weighting method [2] was first presented at the previous IEEE ISIT in San Antonio, Texas. It appears to be an efficient implementation for weighting all the coding distributions (universal over the parameters) corresponding to FSMX models, whose maximum depth does not exceed d. An FSMX model is determined by a proper and complete set S of postfixes. Together these postfixes form a tree that grows towards $-\infty$. Each sequence $x_t$, $t=-\infty$, has a unique postfix in S, i.e. it passes through a unique leaf in the corresponding tree. This postfix (leaf) determines the probability that $x_t$, i.e. the next symbol generated by the (binary) source, is a 1.

The analysis of this weighting method turns out to be very straightforward (see [3]). It shows that the performance is as good as we can possibly hope, not only asymptotically but also for finite sequence lengths. Here we will propose two extensions to the basic weighting method.

II. Coding without knowledge of past symbols

In its basic form the context tree weighting method needs, for every processed symbol $x_t$, $t=1, 2, \ldots, T$, its context $x_{t-1}x_{t-2} \ldots x_{t-d} = \ldots$, where d is the depth of the context tree. This implies that for processing $x_t$, we must have access to $x_{t-d}, x_{t-d+1}, \ldots, x_{t-1}$, and this is not the case, the straightforward way is to start processing only after having seen a full context, i.e. to start processing with $x_{t-d}$. The first context $x_{t-1} x_{t-2} \ldots x_{t-d}$ is send to the decoder in an uncoded way, requiring $d$ bits.

It shows that the performance is as good as we can possibly hope, not only asymptotically but also for finite sequence lengths. Here we will propose two extensions to the basic weighting method.

III. Infinite depth context tree methods

A second unpleasant property of the basic context tree weighting method is that the depth d of the context tree is finite. Only for models that fit into this finite context tree, the weighting method achieves its desirable redundancy bounds.

Our second result is that we can generalize the basic context tree weighting method to the situation where the context tree depth is infinite, and still achieve a storage complexity which is not larger than linear in the sequence length T. The first observation that leads to this result is that we have T contexts each starting (ending) with a (semi-infinite) sequence of 0's, which therefore all differ from each other. The second observation is that the context tree need only contain non-unique nodes that split, or nodes that are unique with a non-unique parent (a node is said to be unique if the subsequence to which it corresponds does only occur once in $x_1 x_2 \ldots x_T$), a node splits if it has more than one child.

It can now be shown that our modified context tree does not contain more than T (internal) nodes. For processing $x_t$, it is however necessary that the sequence $x_{t-1} x_{t-2} \ldots x_{t-d}$ is available in memory. This results in a total complexity not larger than linear in T. Note that this relates to the concept DAWG proposed by Blumer et al. [1].

The performance is as expected, the model redundancy being upperbounded by $|S| - 1$ bits, now for all $S$.

IV. Achieving entropy for arbitrary stationary ergodic sources

Now that we can use the context tree weighting algorithm without knowing past symbols and for infinite depth context trees, we can show that this method achieves entropy for arbitrary stationary and ergodic sources. Let $c(x^T, S) = \log \sum_{z^T} P_z^{(S)}$, the upper bound for the total parameter, model, starting and coding redundancy for large T. Take $S = \{0, 1\}^d$ for arbitrary d, then

$$L(x^T) \leq \sum_{i \in 0} \log \frac{a_i + b_i}{a_i + b_i} \log \frac{a_i + b_i}{a_i + b_i} + c(x^T, S)$$

where $s_i$ and $s_i$ are the number of occurrences of $s_0$ resp. $s_1$ in $x_i^T$, and $\pi_t = P_c(X_{t+d} = 1|X_i^T = s)$, i.e. the actual probability for $s \in S$. Using the ergodic theorem we obtain (entropy in bits) that

$$L(x^T)/T \leq H(X_{t+d}|X_1 \ldots X_d) \quad \text{with probability 1}$$

for $T \to \infty$ for all $d$. Since $\lim_{t \to \infty} H(X_{t+d}|X_1 \ldots X_d) = H_\infty(X)$ it follows, that for stationary ergodic sources the codeword-length $L(x^T)$ divided by the sequence length is, with probability one, not larger than the entropy of the source.

References

