CCIT Report #609
December 2006

On Finite Memory Universal Data Compression and Classification of Individual Sequences

Jacob Ziv
Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel

November 19, 2006

Abstract

Consider the case where consecutive blocks of N letters of a semi-infinite individual sequence $\mathbf{X}$ over a finite-alphabet are being compressed into binary sequences by some one-to-one mapping. No a-priori information about $\mathbf{X}$ is available at the encoder, which must therefore adopt a universal data-compression algorithm.

It is known that if the universal LZ77 data compression algorithm is successively applied to N-blocks then the best error-free compression, for the particular individual sequence $\mathbf{X}$ is achieved as N tends to infinity.

The best possible compression that may be achieved by *any* universal data compression algorithm for *finite* N-blocks is discussed. It is demonstrated that context tree coding essentially achieves it.

Next, consider a device called *classifier* (or discriminator) that observes an individual *training* sequence $\mathbf{X}$. The classifier's task is to examine individual test sequences of length N and decide whether the test N-sequence has the same features as those that are captured by the training sequence $\mathbf{X}$, or is sufficiently different, according to some appropriate criterion. Here again, it is demonstrated that a particular universal context classifier with a storage-space complexity that is linear in N, is essentially optimal. This may contribute a theoretical "individual sequence" justification for the Probabilistic Suffix Tree (PST) approach in learning theory and in computational biology.

Index Terms: Data compression, universal compression, universal classification, context-tree coding.