LDE and a simplified compression-based homogeneity test

Mikhail Malyutov, Northeastern University

Abstract

We continue studying nonparametric compression - based homogeneity tests initiated in [2] under conditions in [2]. The Ziv's test studied in [2] uses computationally intensive evaluation of empirical entropy rate and has the same Large Deviation Exponent (LDE) as the Maximum Likelihood in homogeneity testing, if the size of the training test exceeds constant times the size of the query test. No analysis of the Ziv's test performance in the 'normal range' was done. To analyze strings approximated by Markov Chains of high order d (d-MC) and limited length of training string, applying Ziv's test is problematic. Even more so is its application on line or for simultaneous testing thousands of parallel strings.

The CCC test of [1] using average of several compressed concatenated string increments is computation-elementary and feasible in applications listed above. Its identity of LDE with that of the Ziv's test is established under the assumption that the length of the query test is such that the increments of the compressed concatenated string differ infinitesimally in Probability from those with infinite training string.

References

[1] Malyutov, M. The MDL-principle in attributing authorship of literary texts, *WITMSE 09 International Conference*, Tampere University of Technology, Finland, August 2009, 6 pages, available via IEEXplore.

[2] Ziv, J. On classification and universal data compression. *IEEE Trans. on Inform. Th.*, **34:2**, 278-286, 1988.