

The MDL-PRINCIPLE in ATTRIBUTING AUTHORSHIP of TEXTS

Mikhail Malyutov

Math. Dept., Northeastern University, Boston, MA 02115

ABSTRACT

We study a new *context-free* computationally simple stylometry-based attributor: the *sliced conditional compression complexity (SCCC) of literary texts* which is inspired by the incomputable Kolmogorov conditional complexity. Whereas other stylometry tools can occasionally almost coincide for different authors, our CCC-attributor introduced in Malyutov (2005) is asymptotically strictly minimal for the true author, if the query texts are sufficiently large but much less than the training texts, universal compressor is good and sampling bias is avoided. This classifier simplifies the Ryabko and Astola (2006) homogeneity test (partly based on compression) under insignificant difference of unconditional complexities of training and query texts which can be verified using its asymptotic normality proved in Szpankowski (2001) and elsewhere for IID and Markov sources and normal plots for real literary texts. It is *consistent* under large text approximation as a *stationary ergodic sequence* which follows from the *lower bound for the minimax compression redundancy of piecewise stationary strings* (Merhav (1993)) and from our elementary combinatorial arguments and simulation for IID sources. The SCCC is based on *t-ratio* measuring how many standard deviations are in the mean difference of slices' CCC. The attribution *P-value* can be evaluated based on slices' CCC *asymptotic normality* which is *empirically verified by their normal plots in all cases studied*. The attribution significance statistical evaluation is the main advantage of the SCCC over all previous attributors based on compression complexity.

The *asymptotic SCCC study* is complemented by our attributing the Federalist papers (Madison vs. Hamilton) agreeing with previous results obtained with various classifiers, we also showed a significant (beyond any doubt) mean SCCC-difference between two translations of Shakespeare sonnets into Russian and between the two parts of M. Sholokhov's early short novel and discovered intriguing SCCC-relations between certain Elizabethan poems. At the same time, two different S. Brodsky's novels *deliberately written in different styles* showed insignificant mean CCC-difference.