M. M a l y u t o v (Boston, Northeastern University, USA)**. The minimal description length principle in attributing authorship of texts: a review.**

We study a new *context-free* computationally simple stylometry-based attributor: the *sliced conditional compression complexity* (SCCC) *of literary texts* inspired by the incomputable Kolmogorov conditional complexity idea partially implemented by the so-called **universal compressors** (UC) which adapt to an *unknown* stationary ergodic distribution (SED) of strings attaining asymptotically the Shannon entropy lower bound. $\mathbf{P}$ is the class of SED sources approximated by n-MC's. Compressor family $\mathbf{L} = \{L_n\colon \mathbf{B}^n \to \mathbf{B}^\infty, n = 1, 2, \ldots\}$ is (weakly) *universal*, if for any $P \in \mathbf{P}$ and $\varepsilon > 0$, $\mathbf{B} = \{0, 1\}$), it holds:

$$\lim_{n\to\infty} \mathbf{P}\{x \in \mathbf{B}^n\colon\ |L_n(x)| + \log P(x) \leqslant n\varepsilon\} = 1, \tag{1}$$

where $|L(x)|$ is the length of $L(x)$ and $|L_n(x)| + \log P(x)$ is called *individual redundancy*. Thus for a string generated by a SED, the *UC-compression length is asymptotically its negative loglikelihood* which is used in *nonparametric* statistical inference, if the *likelihood cannot be evaluated analytically*, in particular, for literary texts. Other stylometry tools can occasionally almost coincide for different authors, our CCC-attributor introduced in [1] 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 homogeneity test (partly based on compression) in [3] *under insignificant difference of unconditional complexities of training and query texts* which can be verified using its asymptotic normality proved in [4] for IID and Markov sources and by normal plots for literary case studies. SCCC 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* [2] 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. This enables evaluation of the P-value of statistical significance based on slices' CCC *asymptotic normality* (empirically verified by their normal plots in all cases studied and expected to be proved soon for simplified statistical models of literary texts).

The *asymptotic SCCC study* is complemented by many literary case studies processed by my students and collaborators (Sufeng Li, Irosha Wickramasinghe, Slava Brodsky, Gabriel Cunningham and Andrew Michaelson): attributing the Federalist papers agreeing with previous results, significant (beyond any doubt) mean SCCC-difference between two translations of Shakespeare sonnets into Russian, between the two parts of M. Sholokhov's early short novel and less so between the two Isaiah books from the Bible, intriguing SCCC-relations between certain Elizabethan poems. On the contrary, two different S. Brodsky's novels *deliberately written in different styles* showed insignificant mean CCC-difference.

СПИСОК ЛИТЕРАТУРЫ

1. *Малютов М. Б.* Обзор методов и примеров атрибуции текстов. — Обозрение прикл. и промышл. матем., 2005, т. 12, в. 1, с. 41–77.
2. *Merhav N.* The MDL principle for piecewise stationary sources. — IEEE Trans. Inform. Theory, 1993, v. 39, № 6, p. 1962–1967.
3. *Ryabko B., Astola Y.* Universal codes as a basis for time series testing. — Statist. Methodology, 2006, v. 3, p. 375–397.
4. *Szpankowski W.* Average Case Analysis of Algorithms on Sequences. N. Y.: Wiley, 2001.