Higher Criticism as an Unsupervised Authorship Discriminator

Notebook for PAN at CLEF 2020

Alon Kipnis

Department of Statistics
Stanford University
kipnisal@stanford.edu

Abstract We adapt the Higher Criticism (HC) as an unsupervised untrained discriminator of two documents. Our method takes word-by-word p-values based on a binomial allocation model of words between the documents and combines these p-values to a single test statistic using HC. Large values of HC provide evidence that the two documents are different in terms of authorship. Despite its simplicity, the method achieves competitive results in the Cross-domain Authorship Verification challenge.

1 Overview

Consider two word-frequency tables representing word occurrences across two documents. We would like to check whether the two tables can be regarded as two samples from the same unspecified distribution, or not. Of course, 'word' here could actually assume an extended meaning, such as 'n-grams', n-tuples of consecutive 'dictionary words', or indeed other features of the text that we can render into counts-of-occurrence tables.

In our recent work [4], we proposed a test for this problem based on the Higher Criticism (HC) statistics [1]. This test has two steps. In the first step we perform many exact binomial tests; one test for each word in a prescribed dictionary. The result of each test is a P-value according to a binomial allocation model of the words between the two documents. This model states that each occurrence of the token is equally likely to appear in either document, only accounting for the differences in the sizes of the documents. The second step takes the P-values resulting from the first step and combines them to a single score using the HC statistic.

In a recent study, we show that the method has some optimal theoretical properties when using the binomial word allocation model under 'rare/weak departures' setting [2]. In this setting, if the two distributions are different, they differ only in relatively few words and only by relatively subtle amounts.

In practice, it is unreasonable to assume that the underlying binomial word allocation model is correct, as there may be departures from this model due to topic structure and other violations of binomial word allocation. Nevertheless, the analysis of [4] shows that our method performs well in several authorship challenges even in the presence of such violations. The performance of our method in the PAN2020 Authorship Verification Challenge [3] shows that it serves as an effective authorship discriminator that requires very little tuning.

The section below provides details on the implementation of the method in the aforementioned authorship verification task.

2 Detailed Description

2.1 Vocabulary

We use a vocabulary consisting of 350 of the most common words, bi-grams, and tri-grams in the *small calibration set*. Each document is reduced to its associated word-frequency table over this vocabulary.

Out of the 350 words HC automatically selects a much smaller list of words tailored to each case where it is applied; the evidence for effective discrimination is thought to lie within that selected list [4]. Consequently, the accuracy of the method appears to be insensitive to vocabulary sizes larger than 350.

2.2 Calibration

Our method only requires calibration of the HC score to produce the probability of the event 'same author'. This calibration is done by evaluating the empirical distribution of HC associated with document-pairs over the provided calibration set. We considered the empirical distribution under the cases of 'same author' (H_0) and 'different author' (H_1) separately. For simplicity, we fit a normal distribution to each of these empirical distributions and only store the parameters (μ_i, σ_i^2), $i = 0, 1$ of the fitted distributions.

2.3 Testing

Given a test case (D_1, D_2), we first evaluate the HC score $\mathrm{HC}(D_1, D_2)$ of its two frequency tables. We report on $p(D_1, D_2) = \Pr\left\{\mathrm{HC}(D_1, D_2) | H_0\right\}$ under the assumption that

$$\mathrm{HC}(D_1, D_2) \sim \begin{cases} \mathcal{N}(\mu_0, \sigma_0^2), & H_0, \\ \mathcal{N}(\mu_1, \sigma_1^2), & H_1, \end{cases}$$

where we assume that a priori, the cases H_0 and H_1 are equally likely. If $p(D_1, D_2)$ happens to fall in the interval $(0.45, 0.55)$, we report 0.5 instead.

References

1. Donoho, D., Jin, J.: Higher criticism for detecting sparse heterogeneous mixtures. The Annals of Statistics **32**(3), 962–994 (2004)

2. Donoho, D., Kipnis, A.: Two-sample testing for large, sparse high-dimensional multinomials under rare/weak perturbations (2020), https://arxiv.org/abs/2007.01958
3. Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., Potthast, M., Stein, B.: Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
4. Kipnis, A.: Higher criticism for discriminating word-frequency tables and testing authorship (2019), https://arxiv.org/abs/1911.01208