N-GRAM-BASED TEXT ATTRIBUTION

PATRICK LEAHY

1. Introduction

Quantitative authorship attribution refers to the task of identifying the author of a text based on measurable features of the author's style—a problem that has practical application in areas as diverse as literary scholarship, plagiarism detection, and criminal forensics. Attribution methods generally follow a generative approach, wherein a statistical "profile" is created for a set of candidate authors, based on certain features of the authors' writings, and the author whose profile most closely resembles the corresponding features of the unclassified text is selected. Potential features include word and sentence lengths, letter frequencies, word frequencies, vocabulary richness, word collocations, and more sophisticated (but not necessarily more useful) patterns that appear after syntactic tagging.

Keselj et al. (2003) make a case for the use of character n-grams, i.e. sequences of n characters that occur in the text. One of the most attractive features of this approach is its ease of application: not only does it require no preprocessing, but it is language independent, and can be applied as easily to Chinese or Thai as to English. The particular algorithm they propose relies on a formula for the divergence between two distributions. In this paper we study the effectiveness of a slightly different implementation, one that uses a naïve Bayes classifier. Due to time limitations, we considered only cases in which there are two candidate authors, but our procedure could just as easily be applied to sets of three or more candidates.

2. Methodology

For each candidate author, we created two training texts, each a compilation of multiple texts written by that author. Our program scanned through these training texts and counted the number of occurrences of each character n-gram, ignoring whitespaces but including punctuation. It then compared the two lists and discarded any sequences that were not found in both. This was done to ensure that very infrequent n-grams, such as those that appear only once in the entire training set, would not be included, nor would n-grams that appear very often in one of the author's texts but not at all in others. (The 7-gram "Bingley," for example, appears very often in *Pride and Prejudice*, but not at all in Jane Austen's other novels; thus we would not say that it is a feature indicative of her style.) The counts of the remaining n-grams were then converted into frequencies, which constituted our feature set for that particular author.

Having constructed a profile for each candidate, we then generated n-gram counts for our test document. In order to determine its authorship, we applied a naïve Bayes classifier, the reasoning for which we will describe briefly. By Bayes' rule, the probability that a text belongs to class c, given that it contains features $f_1, f_2, ..., f_n$, is given by

$$p(C = c|f_1, f_2, ..., f_n) = \frac{p(C = c)p(f_1, f_2, ..., f_n|C = c)}{p(f_1, f_2, ..., f_n)}$$

The naïve Bayes model makes the (rather strong, but apparently viable) assumption that each of these features is conditionally independent of all the others, in which case the equation reduces to

$$p(C = c|f_1, f_2, ..., f_n) = \frac{p(C = c)\prod_{i=1}^{n} p(f_i|C = c)}{p(f_1, f_2, ..., f_n)}$$

Since we are not interested in the actual value of $p(C = c|f_1, f_2, ..., f_n)$, only in the likelihood that c is the author relative to the other possibilities, the denominator of this equation can be ignored. Our problem becomes that of finding

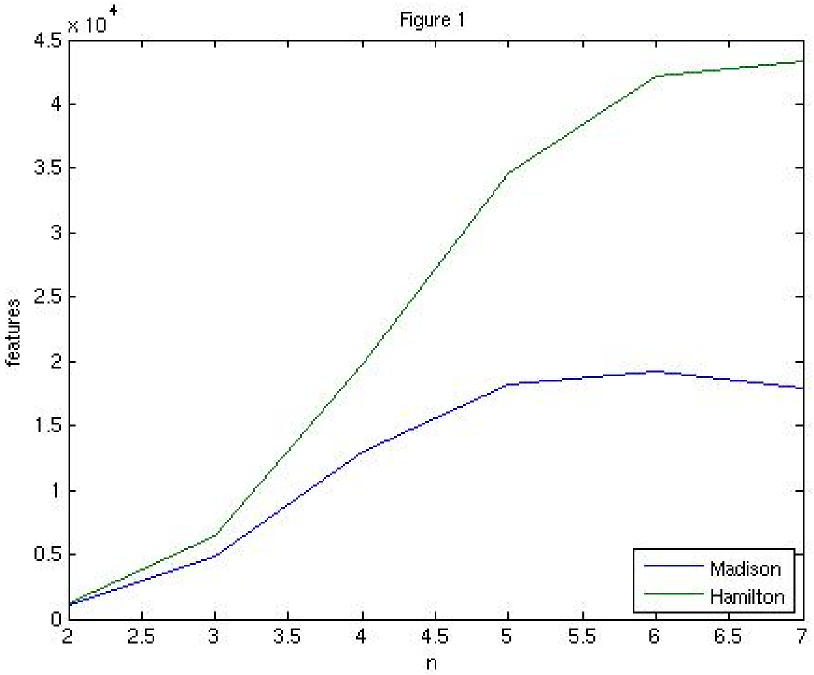
$$\begin{aligned} c_{map} &= \arg\max_c p(C = c)\prod_{i=1}^{n} p(f_i|C = c) \\ &= \arg\max_c[\log p(C = c) + \sum_{i=1}^{n} \log p(f_i|C = c)] \end{aligned}$$

, where the transformation in the second step is applied so. The class priors $p(C = c)$ indicate the prior probability that a text belongs to class c, i.e. was written by candidate author c. Normally, we would estimate these probabilities based on the relative frequency with which each class occurs in the training data, but in our case the training data are the product of artificial selection rather than random sampling; thus we assign each class an equal prior probability, because we have no reason to believe that a test document is more likely to have been written by one candidate author than another. The probability $p(f_i|C = c)$ can be interpreted as the likelihood that a feature f_i occurs in a text given that it was written by candidate author c, and we estimate it by the frequency of that feature in the author's profile. Note that if f_i does not appear anywhere in their profile, then $p(f_i|C = c) = 0$. While there is, in actuality, a non-zero chance that an author will use any particular n-gram, we leave these probabilities as zero given that the appearance of f_i does not *increase* our belief that c is the author of the text.

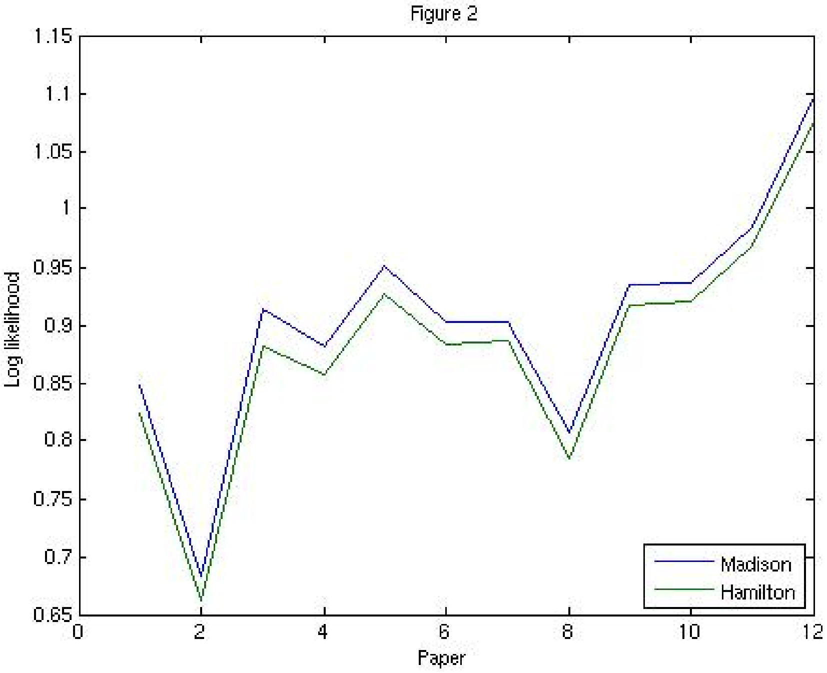
3. Evaluation

We first tested our program on the classic problem of the *Federalist Papers*. Of the 85 papers, which were published under the pseudonym Publius, 51 are known to be the work of Alexander Hamilton, 17 of James Madison, and 5 of John Jay, while 12 papers have been attributed to both Hamilton and Madison. Mosteller and Wallace (1964) were the first to employ statistical techniques to determine the authorship of the 12 "disputed" papers, basing their analysis on the relative frequencies of 30 function words—words, like prepositions, auxiliary verbs, and articles, whose role is grammatical rather than lexical—in the writings of Hamilton and Madison. They concluded that Madison was the author of all 12, and subsequent studies have affirmed their results. The *Federalist Papers* have since become a benchmark of sorts for the evaluation of authorship attribution methods (Holmes and Forsyth, 1995).

We constructed two training texts for both Madison and Hamilton, each comprised of half the papers the respective Founding Father is known to have written. We then ran our program on each of the 12 disputed papers for $n = 2, 3, ..., 7$. Figure 1 shows, for each value of n, the number of useful n-grams in the authors' profiles:



The results were encouraging. Our program identified Madison as the author of all 12 papers, and did so regardless of the value of n. As an example, the results for $n = 3$ are shown in Figure 2 below.



We also ran a second, slightly more extensive test in which the classifier had to discriminate between random excerpts from the novels of Joseph Conrad and Henry James. Each of the 30 test documents—15 by each author—consisted of approximately 1-2,000 words, while the classifier was trained on random selections of approximately 7-8,000 words. The results of this second test are given in the table below.

n	accuracy
2	93.3
3	93.3
4	93.3
5	93.3
6	90.0
7	80.0

Once again, the classifier performed quite well, though its accuracy began to decrease as n increased.

4. Conclusions

Our preliminary results suggest that a naïve Bayes classifier trained on character n-grams can successfully discriminate between authors with fairly similar styles. More generally, character n-grams seem to offer an adequate representation of an author's stylistic "signature." That said, quite a bit of further testing needs to be done. In particular, we would like to determine a) roughly how large the training corpus needs to be for this approach to produce reliable results, b) roughly how large test documents need to be for them to be classified correctly, and c) whether this approach can be applied to more complicated attribution tasks—such as when there are not just 2 but 10 or even 100 candidate authors. We are also interested in whether other methods, such as support vector machines, could perform better than a naïve Bayes classifier.

5. Bibliography

1. Holmes, D., and Forsyth, R. (1995). "The *Federalist* Revisited: New Directions in Authorship Attribution." *Literary and Linguistic Computing*, 10(2):111-127.

2. Keselj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based Author Profiles for Authorship Attribution. *Pacific Association for Computational Linguistics.*

3. Mosteller, F., and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist.* Reading, Massachusetts: Addison-Wesley.