

Authorship Detection Using Support Vector Machines: CS281B Project

Kofi Boakye

May 17, 2004

Abstract

An approach to the task of authorship detection using a support vector machine (SVM) is presented. The dataset used for the experiments is *The Federalist Papers*—a standard in the authorship community. A comparison is made to a word bigram language model (LM) method from which the SVM features arise. Results show that both methods can perform well, though the SVM is more sensitive to unbalanced data as well as the number of bigrams utilized as features.

1 Introduction

Authorship detection—the determination of whether an individual was the author of a given text—has long been of interest in the scholarly communities; tests have been performed on such influential texts as the works of Shakespeare and the Bible. With the dominance of the Internet and email as avenues through which one can engage in economic and intellectual activities, establishing the authenticity of documents is of increasing importance. The techniques used in authorship detection will be of great value in applications such as plagiarism analysis and cybercrime investigation in addition to the traditional uses in analysis of disputed literary and historical documents.

That these documents exist in electronic form facilitates the use of various computational methods, primarily of a statistical nature, to perform detection. These methods seek to capture an author's style, and as such are referred to as *stylometry*. They are based on the assumption that each author's style has certain features that are inaccessible to conscious manipulation either wholly or at least without great difficulty. Some features that have been utilized in stylometric analysis are word lengths, sentence lengths, vocabulary diversity, and word counts[1].

In this paper I describe a project involving the use of a support vector machine (SVM) to perform authorship detection. The performance of the classifier will be compared to that of a language model (LM) system for the same task. The language model is based on word bigram frequencies and these are the features then used by the SVM. The documents used for testing are the articles of *The Federalist Papers*, a corpus which has received much attention in the authorship community.

In the next section I describe the corpus in greater detail. The following section discusses the two approaches mentioned above. Thereafter, performance of the systems is analyzed. Conclusions as well as future work are described in the final section.

2 The Data

The text data used to perform the experiments originates from *The Federalist Papers*, a series of articles originally written under the pseudonym “Publius” by Alexander Hamilton, James Madison, and John Jay from between 1787 and 1788. The articles were written to persuade the citizens of the state of New York to ratify the U.S. Constitution and 77 of them, about 900 to 3500 words in length, appeared in newspapers in the area. These essays, along with an additional eight on the same subject were published in book form in 1788.

The identities of the authors were later assigned to the various articles, but some remain disputed. Of the 85 articles, it is believed that 51 were written by Hamilton, 15 were written by Madison, 5 were written by Jay, and 3 were jointly written by Hamilton and Madison. The authorship of 11 of the articles is disputed, though scholarly consensus generally attributes them all to Madison.

The Federalist Papers are a good experimental dataset for a number of reasons. The principle reason is their frequent usage in authorship experiments. Related to this is the fact that a substantial number of documents exist for the authors (particularly Hamilton and Madison) and they are of moderate length. In addition the text topics are limited in scope and the authors are not only contemporaries, but colleagues. Lastly, though the authorship of some of the texts is disputed, there exists a general consensus label to which one can compare.

3 Approaches

3.1 Language model

The language model approach used for the project is similar to that used by Doddington in [2] for the task of speaker recognition. In the paper, Doddington performs what is essentially authorship detection where the “documents” under investigation are the text output of the transcription of speaker utterances. To

make a decision using this system, a conventional log likelihood ratio test is used. A bigram model for each author is generated using the relative frequencies of the top 2000 word bigrams of the training set. The log likelihood ratio score is then defined to be the log of the ratio of the likelihood of the first author’s model (where model labels are arbitrarily assigned) given the test data to the second author’s model given the same data for a bigram token j , averaged over the bigram tokens of interest (i.e., those among the selected 2000) in the test document. For efficiency, the log likelihood ratio is in reality computed only once for each bigram of type k and then multiplied by the number of occurrences of that bigram in the test text. This gives:

$$\text{Score} = \frac{\sum_k \{N_{tokens}(k) \times \log[\Lambda_{auth1}(k)/\Lambda_{auth2}(k)]\}}{\sum_k \{N_{tokens}(k)\}} \quad (1)$$

where $N_{tokens}(k)$ is the frequency of bigram type k in the test text. For smoothing of the log likelihoods, a value of 0.001 is added to each likelihood before taking the logarithm. For the final decision, the text is attributed to the first author if the score is above zero and the second author otherwise.

3.2 Support vector machine

In the SVM approach, one uses labeled training data—designated as either a positive (+1) or negative (-1) training example depending on the class (here, author) to which it belongs—and attempts to find a separating hyperplane that maximizes the interclass distance. In the common case where the data is not linearly separable, the objective is to find a hyperplane that maximizes the margin while minimizing the penalty due to data points lying on the wrong side. Transformations to the input features may additionally take place before the optimization by way of applying a nonlinear kernel function, but simple linear kernels have shown good results on text data [3] and as such a linear kernel was used for these experiments.

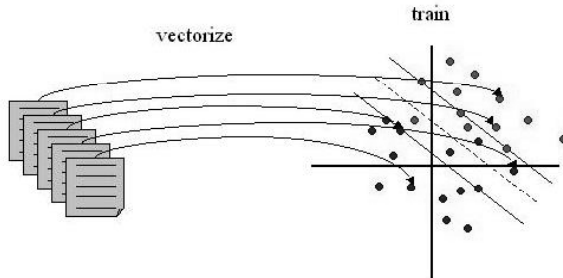


Figure 1: SVM classification of documents.

The features used for the SVM as previously mentioned were the relative frequencies of the top 2000 word bigrams found in the training data. Each document, then, represents a single data point for the SVM classifier as indicated in figure 1. As logarithms of probabilities were not taken to derive these features, the smoothing parameter mentioned in section 3.2 was not applied. The SVM software used for the experiments was SVM^{light}, a publicly available program for building SVM classifiers.

4 Performance

In the tests below the authorship detection task was performed on the disputed Federalist papers using Hamilton and Madison as the possible authors. The following experiments and their outcomes serve to reveal the performance of the SVM system relative to the LM system for a variety of configurations.

Experiment 1

In this experiment, all of the data available (15 Madison texts and 51 Hamilton texts) was used for training. The language model system classifies all test documents as having been written by Madison while the SVM misclassifies all documents as being of Hamilton.

Experiment 2

The results of the first experiment prompted a modification of the SVM training parameters. Specifically, for this experiment the cost of the misclassifications was different for the two classes and was weighted by the class priors as determined by the number of training samples. Again the LM system naturally classifies all documents as being of Madison (since this system was altered in no way) and the SVM also does the same. Closer inspection of the SVM, however, reveals that all 51 of the Hamilton training examples are misclassified; all data points lie to one side of the hyperplane. This, of course, suggests very poor generalization and an overcorrection of the system.

Experiment 3

For this experiment the data was balanced to 15 texts of each author. The experiment was repeated 10 times with random sampling of the 15 Hamilton texts from the 51 total each time. In this case both the LM and SVM systems correctly classify all documents as having been written by Madison.

Experiment 4

In this experiment, the procedure for Experiment 3 is repeated for varying numbers of word bigrams, from 2000 to 50 decrementing by amounts of 50. The average number of misclassifications for each bigram list count was compared for the LM and SVM systems and can be seen in figure 2. As seen in the figure, the language model system appears to outperform the support vector machine

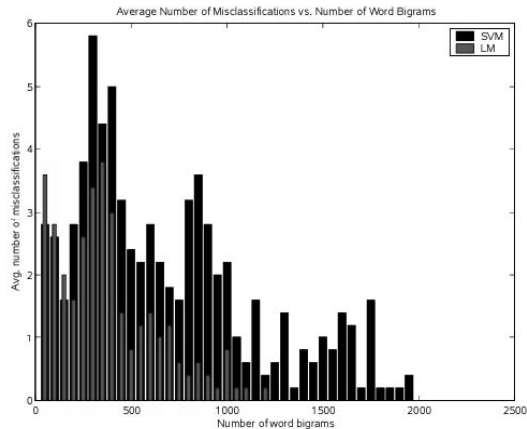


Figure 2: Comparison of classification error for varying number of word bigrams.

system for all but the smallest number of bigrams.

From these results it appears that support vector machines can work effectively in this domain given sufficient tuning. As such the language model system can be seen to be more robust, specifically to phenomena such as unbalanced data and a reduction in the number of bigrams.

5 Conclusions and Future Work

In this paper, an approach to authorship detection using a support vector machine was presented. From an analysis of the results of some select experiments, it seems the SVM lacks robustness to unbalanced data and a reduction in the number of word bigram features, particularly when compared to a language model system. As the SVM requires tuning of parameters, it may be the case that better tuning may mitigate such effects. For example, the regularization parameter, C , was set to 1.0 for all experiments—a common practice, but potentially inappropriate. A more principled approach relying on tuning via cross-validation of held out data would be of interest, though this would reduce the available training data, which is presently small, particularly for the Madison texts. Working on a larger dataset—both in terms of authors and texts—is, then, a natural extension to this work. Lastly, though SVM systems of this type can use a large number of features, it may be of interest to explore feature selection to determine which bigrams are the most discriminative. Results from such an investigation may provide useful information for the area of stylometric analysis of texts.

References

- [1] J. Diedrich, J. Kindermann, E. Leopold, G. Paass, "Author Attribution with Support Vector Machines," *Applied Intelligence*, 2000.
- [2] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," *Proc. Eurospeech'01*, vol. 4, pp. 2521-2524, 2001.
- [3] W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones, and T.R. Leek, "Phonetic Speaker Recognition with Support Vector Machines," *NIPS 2003*.
- [4] G. Fung, "The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization," *Tapia '03*, pp. 42-46, 2003.