Authorship Verification with Compression Features

Cor J. Veenman[2] and Zhenshi Li[1]

[1] Knowledge and Expertise Centre for Intelligent Data-Analysis
Digital Technology and Biometrics Department
Netherlands Forensic Institute
c.veenman@nfi.minvenj.nl
[2] Faculty Technology, Policy and Management
Delft University of Technology
zhenshili@gmail.com

1 Introduction

In the PAN 2013 Author Identification task, the problem was to verify whether a unknown document was written by the same author as a small set of given reference documents. The given reference documents were parts of books with approximately 1000 words per document. The number of reference documents varied between 1 and 10. We followed a statistical pattern recognition approach. There are two possible learning paradigms to apply. The unsupervised way is to establish whether or not the unknown document could be from the same distribution as the reference documents by using the reference documents only. In a supervised approach, a learned classifier establishes a boundary between the reference documents and documents written by other authors using labeled examples of both groups.

In either approach a suitable vector representation of the documents is required. Representations of text documents are typically high dimensional, while the number of reference documents is low. Especially is such small sample cases, an unsupervised approach will expectedly lead to mediocre performing recognition models. On the other hand, to have a representative sampling of documents written by all possible authors, as is required for the supervised approach, is nearly impossible. Fortunately, the given reference documents all had similar genre, theme, and date of writing and the same would hold for the final test cases according to the contest descriptions. As a result, we would only need a representative sample of documents for the given language, genre, theme, and date of writing. Therefore, we followed the supervised learning paradigm. Below we elaborate on the different aspects of this learning task, being: selection of the labeled dataset, the representation of the documents, and the model learning.

Because of time limitations, we only participated in the English Authorship Verification tasks.

2 Data collection

We considered the verification problem as a two-class pattern recognition problem. The reference author is the first *target* class and all other authors are in the second *outlier* class. From the reference author, we have a limited number of given documents: 1-10.

From the outlier class, there are no examples given. Therefore, we needed to collect similar documents as given in the training set to obtain information about the writing style of other authors. It was stated in the task description that in the final test, the documents would have the same language and similar genre, theme, and date of writing as the given training samples. Accordingly, we collected documents with these properties.

In order to obtain similar English documents as the ones provided by the PAN 2013 organization, we searched the Internet for substrings of the provided documents. We found several of the training documents in he bookboon.com repository, which provides open access for students to text books. From this repository we selected 66 textbooks from the Engineering (Chemical Engineering, Construction Engineering, Electrical Engineering, Energy Engineering, Environmental Engineering, Mechanical Engineering, Nanotechnology and Petroleum Engineering) and IT & Programming sections. The books were authored by 46 different authors.

We prepared the documents similarly as the training/reference material. That is, we converted the text sections to UTF8 format and split the books into documents of 6,000 to 8,000 characters each. This resulted in 2 to 75 documents per text book with roughly 1,000 words each.

3 Document representation

As document representation, we chose the compression distance to other documents. Several compression distance measures have been proposed in the past [6]. We used the Compression Dissimilarity Measure (CDM) [3]:

$$CDM(\boldsymbol{x}, \boldsymbol{y}) = \frac{C(\boldsymbol{xy})}{C(\boldsymbol{x}) + C(\boldsymbol{y})}, \tag{1}$$

where $C(\boldsymbol{x})$ is the size of the compressed object $\boldsymbol{x}$ and $\boldsymbol{xy}$ is the concatenation of $\boldsymbol{x}$ and $\boldsymbol{y}$. To obtain $C(\boldsymbol{x})$, we used the best text compressor currently available. In such a way, the best approximation of the Kolmogorov complexity of a text string is obtained, which is the underlying theory behind the compression based distance measure [1]. In [2], we showed that better compressors indeed result in better recognition performance. From the text compression benchmark [5], it can be seen that adaptive statistical data compressors using context modeling and prediction are the current best compressors. In particular, we used the Prediction by Partial Matching (PPMd) compressor by [7].

4 Model learning

We submitted three approaches to the English Authorship Verification task, with an increasing modeling complexity. All approaches used the collected documents to represent the outlier class and compression distances as document representation.

4.1 Nearest Neighbor with Compression Distances

The simplest approach we prepared is purely based on compression distances between the unknown document on the one hand and the reference documents and collected documents on the other hand, see Figure 1. The decision rule is as follows: if the closest document in the dataset is from the reference author, we conclude that the unknown document is from that author too. Otherwise, if a document from any author from our collected corpus is closer, we conclude that the unknown document is not written by the reference author. Such a 1-nearest neighbor approach is known to be sensitive for overfitting or outliers in the data. On the other hand, the length of documents probably mitigates this problem to a certain extend. That is, for longer documents it is less probable that accidentally a document of another author is more similar than that of the true author. On the provided cases, this approach resulted in 1 error out of 10 (10%).

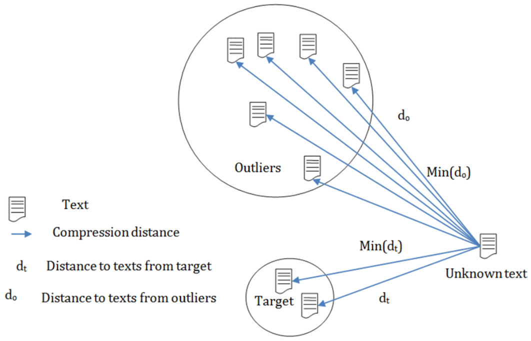

Figure 1. Nearest neighbor with compression distances

4.2 Two-class classification in Compression Prototype Space

For the second approach, we learned a two-class classifier in prototype space as in [4], see Figure 2. The first (target) class is the class of the documents of the reference author and the second (outlier) class is that with all collected documents (by other authors). All documents are represented by their compression distance toward a number of selected documents from the training set, i.e. the prototypes. To be able to compute as much of the compression distances beforehand, the prototypes are selected from the collected documents, i.e. the outlier class. Since this class is diverse, the distances to these documents will give a rich representation. Because of the limited amount of available documents, the prototypes themselves were not removed from the dataset for model learning. Per test case, we learned a classifier with the reference and collected documents and applied the resulting classifier to the unknown document. The classifier decides if the document is from the reference author or not. As classifier, we used the Lowest Error in a Sparse Subspace (LESS) classifier [8] that is able to deal with small sample size problems, see Eq. 2.

$$\min \sum_{j=1}^{p} w_j + C\big(\sum_{i=1}^{n_t} \xi_{ti} + \sum_{i=1}^{n_o} \xi_{oi}\big), \tag{2}$$

$$\text{subject to:} \quad \begin{cases} \boldsymbol{x} \in X_t, & \sum_{j=1}^{p} w_j \phi(\boldsymbol{x}, j) \geq 1 - \xi_{ti} \\ \boldsymbol{x} \in X_o, & \sum_{j=1}^{p} w_j \phi(\boldsymbol{x}, j) < -1 + \xi_{oi} \end{cases}$$

$$\text{where} \quad \phi(\boldsymbol{x}, j) = \big(x_j - \mu_{tj}\big)^2 - \big(x_j - \mu_{oj}\big)^2$$

$$\text{and} \quad w_j \geq 0, \ \xi_i \geq 0.$$

Here, $\boldsymbol{w}$ is the model vector with weights w_j per dimension, X_t and X_o contain the target and outlier samples, n_t and n_o are the sizes of X_t and X_o, μ_t and μ_o are the mean vectors of X_t and X_o, p is the number of dimensions (in this case prototypes), and ξ_{ti} and ξ_{oi} are the slack variables for documents from the respective classes to enable the modeling of inseparable classes. Further, C is a tunable parameter that balances model sparseness against model accuracy.

Our collected dataset is a lot bigger than the provided reference documents. Without precautions, the scarce target class would be neglected by the model, just because it is scarce. From the training cases it turned out that roughly in 50% of the cases the unknown document is from the reference author (target). Therefore, we made both classes equally important in the LESS model in order to deal with these strongly unbalanced classes, see Eq. 3.

$$\min \sum_{j=1}^{p} w_j + C\big(\frac{1}{n_t}\sum_{i=1}^{n_t} \xi_{ti} + \frac{1}{n_o}\sum_{i=1}^{n_o} \xi_{oi}\big), \tag{3}$$

$$\text{subject to:} \quad \begin{cases} \boldsymbol{x} \in X_t, & \sum_{j=1}^{p} w_j \phi(\boldsymbol{x}, j) \geq 1 - \xi_{ti} \\ \boldsymbol{x} \in X_o, & \sum_{j=1}^{p} w_j \phi(\boldsymbol{x}, j) < -1 + \xi_{oi} \end{cases}$$

$$\text{where} \quad \phi(\boldsymbol{x}, j) = \big(x_j - \mu_{tj}\big)^2 - \big(x_j - \mu_{oj}\big)^2$$

$$\text{and} \quad w_j \geq 0, \ \xi_i \geq 0.$$

The number of prototypes and the C trade-off parameter for LESS, we optimized on the collected corpus. In turn, we selected one of the 46 authors as reference author and took up to 10 of his documents as reference documents. We put a selection of documents from the that author and from other authors aside for testing. In this way, we established $p = 200$ and $C = 10,000$.

We ran several experiments with different samples of 200 prototypes on the provided cases. This resulted in 26% average error on the 10 cases.

4.3 Bootstrapped document samples

In the last approach, we attempt to make the method more robust for the low number of document samples in the reference class. Ultimately, there could be only one reference document. Therefore, we resampled the available reference documents as we did in

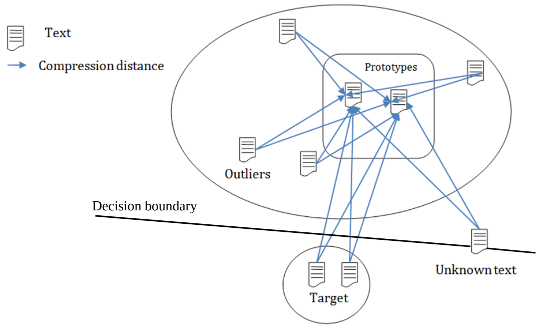

Figure 2. Classifier in prototype space

[2], see Figure 3. To make the method generic, we first merged the available reference documents (1-10) in given order into a single document. Then, we sampled 50 documents from the merged document with the same average size as the given reference documents (± 1000 words). Now, the target class always contains 50 documents and the outlier class all other collected documents. The method proceeds as the previous one. The tuning parameters for the number p of prototypes and LESS trade-off parameter C, were not that sensitive and were kept the same.

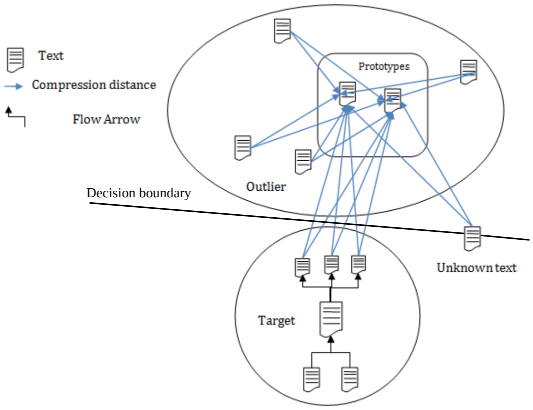

Figure 3. Bootstrapped documents in prototype space

We ran several experiments with different bootstrapped document samples and different prototypes on the provided cases. This resulted in 16% average error on the 10 cases.

5 Conclusions

We prepared three submissions for the PAN 2013 Authorship Verification task. An important part of our approach, was careful selection of documents not authored by the reference author, the outlier class. All our submissions used compression distances as document representation. Two of our submissions additionally used a representation with compression distances to prototypes, that were selected as a subset of the outlier class. The last submission used document resampling to increase the size of the target class. On the provided 10 test cases the submissions had similar performance: 1-2 errors. With only 10 test cases significance of differences in performance could not be established. With these submissions we obtained the best (ex aequo) score $F_1 = 0.80$ out of the 16 teams, that submitted for the English Authorship Verification task.

References

1. Cilibrasi, R., Vitányi, P.M.B.: Clustering by compression. IEEE Transactions on Information Theory 51(4), 1523–1545 (Apr 2005)
2. de Graaff, R., Veenman, C.: Bootstrapped authorship attribution in compression space. In: CLEF 2012 Evaluation Labs: PAN - Author Identification (2012)
3. Keogh, E., Lonardi, S., Ratanamahatana, C.: Towards parameter-free data mining. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 206–215 (2004)
4. Lambers, M., Veenman, C.: Forensic authorship attribution using compression distances to prototypes. In: Proceedings of the Third International Workshop on Computational Forensics, The Hague, The Netherlands, August 13-14. pp. 13–24. Springer-Verlag, Berlin, Heidelberg (2009)
5. Mahoney, M.: Large text compression benchmark, http://www.mattmahoney.net/text/text.html
6. Sculley, D., Brodley, C.E.: Compression and machine learning: A new perspective on feature space vectors. In: Proceedings of the Data Compression Conference. pp. 332–332. DCC ’06, IEEE Computer Society, Washington, DC, USA (2006)
7. Shkarin, D.: PPM: one step to practicality. In: Proceedings of the Data Compression Conference. vol. DDC ’02, p. 202. IEEE Computer Society (2002)
8. Veenman, C., Tax, D.: LESS: a model-based classifier for sparse subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(9), 1496–1500 (Sep 2005)