Bootstrapped Authorship Attribution in Compression Space

Notebook for PAN at CLEF2012

Ramon de Graaff[1] and Cor J. Veenman[2]

[1] Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
ramondegraaff@gmail.com
[2] Digital Technology and Biometrics Department
Netherlands Forensics Institute
c.veenman@nfi.minvenj.nl

Abstract From a machine learning standpoint, the PAN 2012 Lab contest had one major challenge. In all authorship attribution tasks, the number of training documents was extremely low. We extended our previous work, in which compression distances to randomly selected prototype documents from the training corpus were used as feature representation. A supervised multi-class classifier was learned in the resulting feature space using the remaining documents. Inspired by the bootstrapped resampling method, we now drew document samples from the few source documents in order to obtain sufficient prototypes and samples to learn a supervised classifier. Using internal validation, we tuned the size of the document samples, compression method, distance measure, classification method, and decision threshold (open-class tasks) for optimal F_1 score. With this scheme we submitted for the closed-class and open-class author identification tasks. In the overall results for these tasks we achieved a shared fourth ranking, based on the reported average recall of the 11 teams.

1 Introduction

This years PAN 2012 Lab author identification task had two sub-tasks: the traditional Authorship Attribution and Sexual Predator Identification. From the Authorship Attribution sub-task our interest goes to closed-class and open-class (traditional) authorship attribution. The second problem within this sub-task was authorship clustering or intrinsic plagiarism, which we did not consider.

The datasets provided had a very low number of candidate authors compared to the last years contest. Moreover, per author the number of sample documents was very low, i.e., only two documents per author. Third, the size of the sample documents was relatively large: order 10kB - 100kB. From a machine learning standpoint, the first point makes live easier, while the second point is a major challenge. The sample documents that make up the training set for model learning, hardly enable to generalize with two samples per class. The situation is even worse, to be able to do internal validation for model selection, one document should be kept apart, so that only one document remains for training of the recognition models.

In our previous work [9] we proposed the Compression Distance to Prototypes (CDP) method. We applied this method to datasets with similar characteristics as the PAN 2011 Lab authorship attribution contest. These characteristics are a high number of authors, per author tens of sample documents and the size of the training documents was relatively small. In short, the CDP methods randomly selects per author a part of the provided document corpus as prototypes. The remaining documents are used as training set for recognition model learning. The feature representation of the training set is computed as compression distances to the prototypes. Compression distances are distance measures in the sense that similar documents have small distances and dissimilar documents have larger distances.

Without adaptation, our previous work could not be applied for the PAN 2012 Lab, since there is only one training document per author. One possible adaptation is to compute the compression distance from a test document to all training documents and attribute a test document to the author of the closest training document. This 1-nearest-neighbor procedure is known to be sensitive to noise or, in other words, it easily overfits to the training corpus. Besides the 1-nearest neighbor rule, with effectively only one document for model learning there are hardly any methods that can be applied. Moreover, the same risk of overfitting would apply.

Here, we propose an adaptation of our previous work in which we regenerate a training set from the given corpus such that statistical model learning becomes feasible. In the next section, we first pose the problems derived from the PAN 2012 Lab subtask we take part in. Then, we elaborate on our method and proposed extensions. In the following section, we apply our method to the training corpus for parameter tuning using internal cross-validation. Among the model parameters are the compression method for the compression distance computation and the compression distance measure itself. Finally, we describe the results of applying the tuned models to the test corpus as submitted for the contest. We wrap up with concluding remarks about the obtained results.

2 Problem statement

The problems of the traditional authorship attribution sub-task, that we considered for the PAN 2012 Lab, are the closed-class and open-class authorship attribution problems. As statistical pattern recognition problem, closed-class authorship attribution comes down to a standard multi-class classification problem, where each class is one of the known authors. Open-class authorship attribution can be seen as a multi-class problem, where one class is added representing all unknown authors. The problem is to find proper representations and models for the closed-class and open-class authorship attribution task, where precision, recall and F_1 measure will be used as evaluation metrics. These measures are defined as:

Precision P_A for author A is defined as:

$$P_A = \frac{\text{correct(A)}}{\text{retrieved-documents(A)}} \equiv \frac{TP_A}{TP_A + FP_A}, \tag{1}$$

where TP_A (True Positive) is the number of documents that are correctly attributed to author A and FP_A (False Positive) is the number of documents that are incorrectly attributed to author A.

Recall R_A for author A is defined as:

$$R_A = \frac{\text{correct(A)}}{\text{relevant-documents(A)}} \equiv \frac{TP_A}{TP_A + FN_A} \tag{2}$$

where FN_A (False Negative) is the number of missed attributions to author A. The F_1 measure [15] is defined as the harmonic mean of recall and precision:

$$F_1 = 2 \cdot \frac{P_A \cdot R_A}{P_A + R_A} \tag{3}$$

These measures can be aggregated by averaging, either author based or document based, leading to macro and micro averages, respectively [19]. For instance, the macro averaged recall R_{macro} is defined as:

$$R_{macro} = \frac{1}{n} \sum_{i=1}^{n} R_{A_i} \tag{4}$$

and the micro averaged recall R_{micro} is defined as:

$$R_{micro} = \frac{1}{k} \sum_{i=1}^{n} |D_{A_i}| \cdot R_{A_i}, \tag{5}$$

where $|D_{A_i}|$ is the number of documents in the test set for author A_i, and $k = \sum_{i=1}^{n} |D_{A_i}|$.

3 Method

The method we propose for this task is based on the Compression Distance to Prototypes (CDP) method we reported earlier in [9]. We first summarize the CDP method and then extend it with provisions to deal with the extremely small sample size of the contest, i.e., one training sample per author.

The CDP method deserves its name from the way the training documents are represented, i.e., its feature representation. In contrast with typical representations for text documents with lexical, syntactical and structural features, we represent a document as being similar (or dissimilar) to a set of other documents. Such a dissimilarity based representation was proposed earlier in [14] and has proven to give competitive classification results. It can be favorable for obtaining lower dimensional representations, especially if suitable distance measures are available. Importantly, the distance measure to be used should discriminate the samples such that dissimilar samples have large distances and similar samples have small distances. Several compression based distances have these properties and have been applied successfully in different domains [1], [2], [11], [7], [8], [18]. These compression-based approaches are practical implementations of the information distances expressed in the non-computable Kolmogorov complexity [12]. In [9], we applied the Compression Dissimilarity Measure (CDM) [7]:

$$CDM(\boldsymbol{x}, \boldsymbol{y}) = \frac{C(\boldsymbol{xy})}{C(\boldsymbol{x}) + C(\boldsymbol{y})}, \tag{6}$$

where $C(\boldsymbol{x})$ is the size of the compressed object $\boldsymbol{x}$ and $\boldsymbol{xy}$ is the concatenation of $\boldsymbol{x}$ and $\boldsymbol{y}$. Essential in all these measures is a compressor that finds the smallest possible encoding of, in this case, the sample documents. In [9], we used the LZ76 compression method [10]. The contribution of that work was to use compression-based distances as feature representation.

3.1 Bootstrapped document samples

After having defined the representation of the documents, we propose a way of regenerating sample documents from the single given training document per author. This is possible because, fortunately, the documents of the PAN 2012 Lab contest are relatively large. We use the same idea underlying bootstrapping, a well known resampling method for generalization error estimation [6]. The rationale behind the bootstrapped resampling method is the best representation of the data distribution is the given dataset itself. In our case, this translates to: the best representation of the writing style of an author is the one document that we have.

The method works as follows. First we draw a prototype for the given author from the start of the document with a certain length. The length of the prototype is a parameter to select. Then we proceed similar to the bootstrapped resampling method with the remaining part of the document. That is, in order to get training sample documents written by the author, we draw from her 'model' with replacement, where the model is the one source document. The sampling of training samples works as follows. The starting point in the (remaining) document is chosen randomly. Then the required number of characters is read. In case the sample would read over the end of the document, it continues reading at the start of the document until the required length is obtained.

3.2 Classifier learning

Closed-class recognition Finally, a classifier must be learned in the compression distance space. For this purpose any multi-class classifier can be used. For model selection and parameter tuning we use the F_1 measure, which is an aggregation of precision and recall that will be used as performance measures in the contest.

Open-class recognition For the open-class tasks, we additionally had to estimate a threshold for deciding for an unknown author. That is, the classifier decides for the most probable class, unless its probability is below the given threshold. In that case, it decides *none* of the known classes. We estimated the threshold by trying all thresholds in the training set that resulted in the best averaged F_1 score on the test set, where one class was left out in turn and considered as *none* of the known classes.

3.3 Method parameters

Below, we list the parameters involved in the method. Some of these are selected a priori, some are estimated in case they seem to be less dependent on the dataset and others are optimized per dataset as will be described in the Section Experiments.

1. *Distance measure*: Besides the already mentioned CDM in this work we also consider the Normalized Compression Distance (NCD) [2] as dissimilarity measure:

$$NCD(\boldsymbol{x}, \boldsymbol{y}) = \frac{C(\boldsymbol{xy}) - min\{C(\boldsymbol{x}), C(\boldsymbol{y})\}}{max\{C(\boldsymbol{x}), C(\boldsymbol{y})\}} \tag{7}$$

2. *Compressor*: The compressor is the core ingredient of the compression-based distance measures. From theory the best compressor should be used. Therefore, in this work we additionally considered a variant of the PPM algorithm, PPMd, that is among the best for text compression [3], [17], [13].
3. *Bootstrapping*: The bootstrapping method adds four additional parameters: the number of prototypes per author, the number of drawn training samples, the size of the prototypes, and the size of training samples.
4. *Classification*: The classification method to be applied for closed-class and open-class recognition.
5. *Open-class recognition*: the threshold for open-class recognition.

4 Experiments

Below, we first describe the different datasets in the Authorship Attribution sub-task we submitted our runs for. Then, we describe the way we handled the method parameters and we list the internal cross-validation results and submitted runs of the experiments.

4.1 Datasets

The PAN12 authorship identification sub-task had several datasets with different numbers of authors. In every dataset for each author two documents were given.

The mean document sizes and standard deviations for the datasets of PAN12 are shown in Figure 1. Details on the datasets in PAN12 can be found in Table 1.

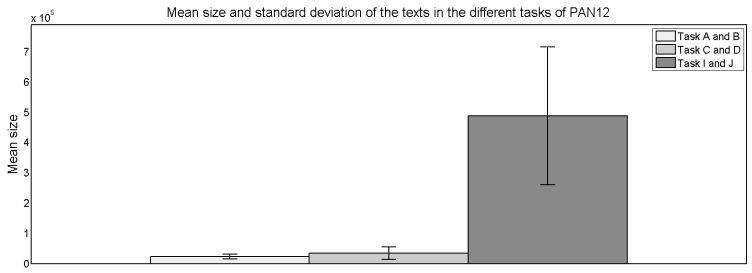

Figure 1. PAN12 Distribution of training corpus of the PAN12

Dataset	No. classes	No. documents	Size (kB)
TASK A TRAIN	3	6	9-32
TASK A TEST	3	6	5-43
TASK B TEST	4	10	10-39
TASK C TRAIN	8	16	11-72
TASK C TEST	8	8	10-43
TASK D TEST	9	17	10-74
TASK I TRAIN	14	28	179-1023
TASK I TEST	14	14	231-1123
TASK J TEST	15	16	98-1271

Table 1. Dataset 2012 The number of authors and number of documents per dataset

4.2 Parameter setting

The parameters of the method were either chosen a priori, estimated globally for all tasks, tuned per task by two-fold cross-validation for optimal averaged F_1 score, or the exploration of parameter was part of the experiments. Two-fold cross-validation was conducted by taking for each author the first document as training document for prototype sampling and bootstrapped sampling and the second for validation. The validation document was divided up in three parts, to enable better F_1 score differentiation between several parameter settings. Then the sampling and validation set was rotated by using the second document for training and the first for validation. This process was repeated 10 times and the results averaged.

1. *Distance measure*: Preliminary experiments on the TASK I dataset showed that the NCD and CDM distance measure performed similarly. The reported experiments were therefore conducted with the CDM measure as before, i.e., as in [9].
2. *Compressor*: Preliminary experiments on the PAN 2011 Lab data and the PAN 2012 Lab TASK I showed that the PPMd compressor clearly outperformed the previously used LZ76 compressor. The reported experiments were therefore conducted with the PPMd method.
3. *Bootstrapping*: Because we draw the prototypes without replacement, we fixed the number of prototypes to one per author. The size of the prototypes, the number and size of training samples are tuned through two-fold cross-validation for optimal F_1 score per task.
4. *Classification*: The classification method is explored in the experiments.
5. *Open-class recognition*: The threshold for open-class attribution is established through two-fold cross-validation for optimal F_1 score. In the averaging of F_1 scores the unknown class is considered equally important as all known authors taken together.

We implemented the method in Matlab and used the pattern recognition toolbox PRTools [5] for the classification models.

4.3 Results

We separate the description of the experiments in the closed-class tasks and the open-class tasks.

4.4 Closed-class

The datasets used in these tasks consist of three, eight and fourteen authors for TASK A, TASK C and TASK I, respectively. The sizes of the training documents for these tasks differ quite a lot as can be seen in Table 1. In Figures 2, 3 and 4, the internal cross-validation results can be seen for some optimized parameter settings as prototype size and bootstrapped training sample document size. Based on these figures we set the method parameters to be used in the runs for submission. We selected Fisher linear discriminant [4] as classifier for all tasks and the number of bootstrapped training samples to thirty. Further, we set the prototype size, the size of the bootstrapped training samples as shown with the figures for the respective tasks (Figures 2, 3 and 4) and Table 2.

For the submissions, we could exploit both training documents for each author, since the method parameters were tuned. For the first submission, we used two prototypes per author. That is, we took a prototype from both training documents of each author. From the remaining part of the training documents, we drew thirty samples of a size conform the tuned parameter specified in Table 2. For the second submission, we used one prototype per author from the first document and sample both documents for trainings samples with the given parameters.

This resulted in models based on more training data than in the internal validation. Expectedly, this could only improve the performance. However, the performance of SUBMISSION 2 is quite worse than the performance of the internal validation and SUBMISSION 1. SUBMISSION 1 performs pretty well. The performances of the two runs on the test documents provided by PAN12 are shown in Table 3.

			Macro			Micro		
Dataset	Prototypes	Samples	Precision	Recall	F1-measure	Precision	Recall	F1-measure
TASK A	20%	70%	0.64	0.71	0.66	0.64	0.71	0.66
TASK C	20%	90%	0.77	0.82	0.78	0.77	0.82	0.78
TASK I	50%	70%	0.81	0.86	0.82	0.81	0.86	0.82

Table 2. Internal validation The internal validation and parameters on the closed-class datasets of PAN12, 10-repeat 2-fold cross validation

		Macro			Micro		
	Dataset	Precision	Recall	F1-measure	Precision	Recall	F1-measure
ONE	Task A	1.0	1.0	1.0	1.0	1.0	1.0
ONE	Task C	0.81	0.88	0.83	0.81	0.88	0.83
ONE	Task I	0.60	0.71	0.63	0.60	0.71	0.63
TWO	Task A	0.50	0.67	0.56	0.50	0.67	0.56
TWO	Task C	0.55	0.79	0.55	0.76	0.47	0.38
TWO	Task I	0.42	0.50	0.44	0.42	0.50	0.44

Table 3. PAN12 Lab (closed-class) The performances on the closed-class tasks of the PAN12 Lab for SUBMISSION 1 and SUBMISSION 2.

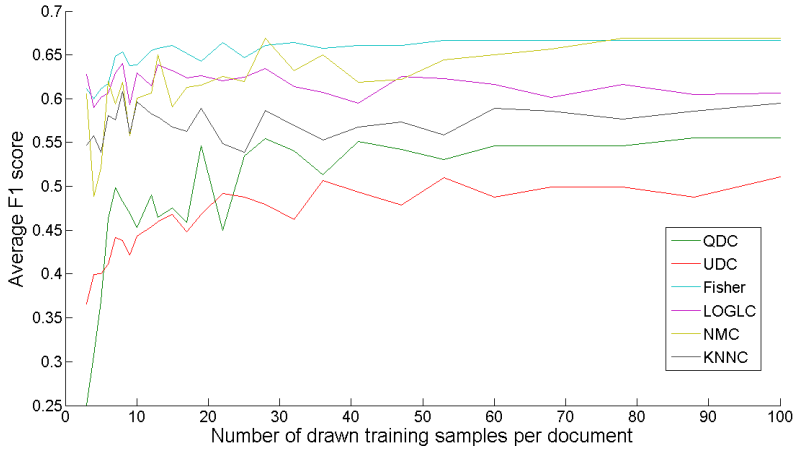

Figure 2. Task A and B: Average F_1 score as a function of the number of bootstrapped training samples using CDM and PPMd. The optimized prototype size is 20% of the document and the optimized bootstrapped sample size is 70% of the document.

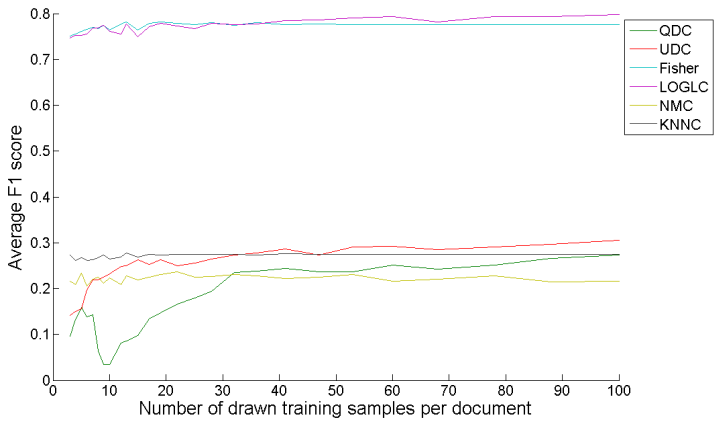

Figure 3. Task C and D: Average F_1 score as a function of the number of bootstrapped training samples using CDM and PPMd. The optimized prototype size is 20% of the document and the optimized bootstrapped sample size is 90% of the document.

4.5 Open-class

The training data for the open-class tasks TASK B, TASK D and TASK J is the same as for the corresponding closed-class tasks TASK A, TASK C and TASK I.

The internal validation on the open datasets is done using a ten repeat experiment on the dataset while measuring the F_1 performance. Per repeat, every author is two times offered as the 'Unknown', each of its training documents once. We compute the

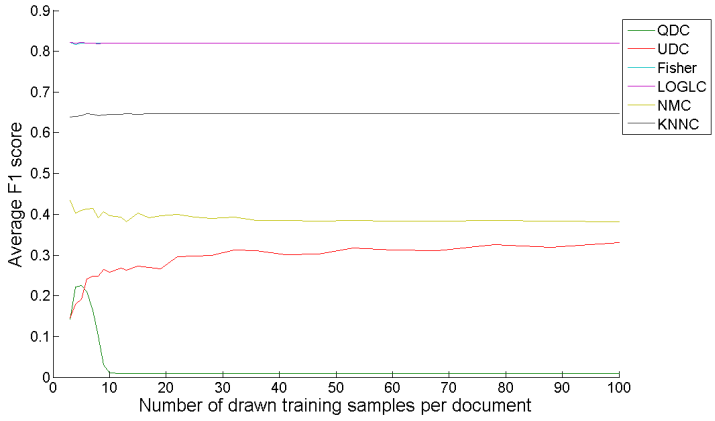

Figure 4. Task I and J: Average F_1 score as a function of the number of bootstrapped training samples using CDM and PPMd. The optimized prototype size is 50% of the document and the optimized bootstrapped sample size is 70% of the document.

F_1 score in two ways, that we denote as PN and $P50$. With PN, we express that the unknown author is as important to recognize as any single author. With $P50$, we express that the unknown author is as important as all remaining authors together. Hence, the unknown author is weighted for 50% and the other authors together as the other 50%.

In Table 5, we see that PN is higher than $P50$ on every dataset. This corresponds to our expectation, because here only $\frac{1}{n}$th, with n authors, is offered as 'Unknown'. Distinguishing the 'Unknown' author is here as important as attributing the test documents to every known author. We introduced $P50$ because we expect more 'Unknown' authors in the testsets provided by PAN12 than only $\frac{1}{n}$th.

In Table 4, the number of known versus unknown authors in the testset provided by PAN12 is shown, as well as the optimized thresholds. In Table 6 the performances are shown for the submissions on the testset provided by PAN12. After we optimized the thresholds, we take the same models for SUBMISSION 1 and SUBMISSION 2 as we did for the closed-class. That is, n prototypes for SUBMISSION 1 and $\frac{1}{2}n$ for SUBMISSION 2 where n is the number of train documents. As we expect more or equal documents of the 'Unknown' author, we submit the models with the threshold $P50_T$. In TASK B, the number of documents by 'Unkown' authors is only four, the threshold PN_T would perform slightly better. Fortunately in TASK D, the number of documents by 'Unknown' authors is nine, which is about half of the dataset. The threshold $P50_T$ performs a lot better than threshold PN_T. In TASK J, both thresholds came up with the same labeling for the test dataset. The models from SUBMISSION 2 came up with the same labels for both thresholds on all tasks. Clearly, for optimal performance the proportion of known and unknown authors should be known beforehand.

Dataset	Known	Unknown	Total	PN_T	$P50_T$
TASK B	6	4	10	0.9749	0.9812
TASK D	8	9	17	$5.45 \cdot 10^{-4}$	0.0084
TASK J	14	2	16	$6.90 \cdot 10^{-5}$	$4.19 \cdot 10^{-4}$

Table 4. PAN12 (open-class) Testset Distribution of the provided test documents of the open-class tasks for PAN12 Lab, including the calculated thresholds

			Macro			Micro		
Dataset	Prototypes	Samples	Precision	Recall	F1-measure	Precision	Recall	F1-measure
TASK B (PN_T)	20%	70%	0.47	0.52	0.48	0.47	0.52	0.48
TASK B ($P50_T$)	20%	70%	0.45	0.51	0.47	0.45	0.51	0.47
TASK D (PN_T)	20%	90%	0.64	0.70	0.64	0.64	0.70	0.64
TASK D ($P50_T$)	20%	90%	0.47	0.51	0.48	0.47	0.51	0.48
TASK J (PN_T)	50%	70%	0.75	0.81	0.77	0.75	0.81	0.77
TASK J ($P50_T$)	50%	70%	0.55	0.65	0.59	0.55	0.65	0.59

Table 5. Internal validation The internal validation performances on open-class datasets of PAN12, 10 repeat 2-fold cross validation

5 Conclusion

For the PAN 2012 Traditional Authorship Attribution tasks, we modified our previous work to deal with the major challenge of the provided datasets. That is, for all tasks the number of training documents per author was only two. To be able to do model selection, one document must be kept apart, so that one document could be exploited for model learning. We proposed a method for generating additional training documents inspired by the bootstrapped resampling method for generalization error estimation. Both the internal validation results and the results of the submitted runs on the test data show, that this resulted in a promising method for closed-class and open-class authorship attribution. In the overall results, we achieved a shared fourth ranking for the authorship attribution tasks, based on the reported average recall of the 11 teams. Further, the CDM compression distance in combination with the PPMd compressor outperformed other combinations, which is in line with results reported in [16].

The open-class experiments showed how important it is that the training data and test dataset have the same characteristics for statistical pattern recognition methods. In this case, the proportion of known and unknown authors could not be derived from the training data. We guessed the unknown authors to be as frequent as all known authors together. Clearly, this assumption has a strong impact on the results. This was shown in experiments in which we assumed that an unknown author to be as frequent as any single known author.

Finally, the improved performance of SUBMISSION 1 over SUBMISSION 2 shows that with more prototypes a better representation of the documents and a better recognition performance can be obtained. This is an aspect that should be explored further. For instance, the number of prototypes could be optimized by exploiting the document bootstrapping for prototypes too.

		Macro			Micro		
	Dataset	Precision	Recall	F1-measure	Precision	Recall	F1-measure
ONE	Task B (PN_T)	0.75	0.63	0.63	0.70	0.60	0.60
	Task B ($P50_T$)	0.48	0.50	0.44	0.46	0.50	0.44
	Task D (PN_T)	0.55	0.79	0.55	0.76	0.47	0.38
	Task D ($P50_T$)	0.71	0.85	0.75	0.78	0.76	0.75
	Task J (PN_T)	0.72	0.80	0.74	0.68	0.75	0.70
	Task J ($P50_T$)	0.72	0.80	0.74	0.68	0.75	0.70
TWO	Task B	0.16	0.31	0.21	0.18	0.30	0.21
	Task D	0.18	0.44	0.22	0.10	0.24	0.12
	Task J	0.62	0.73	0.65	0.58	0.69	0.61

Table 6. PAN12 Lab (open class) The performances on the open class tasks of the PAN12 Lab for SUBMISSION 1 and SUBMISSION 2.

References

1. Benedetto, D., Caglioti, E., Loreto, V.: Language trees and zipping. Phys. Rev. Lett. 88(4), 048702 (Jan 2002)
2. Cilibrasi, R., Vitányi, P.M.B.: Clustering by compression. IEEE Transactions on Information Theory 51(4), 1523–1545 (Apr 2005)
3. Cleary, J., Witten, I.: Data compression using adaptive coding and partial string matching. IEEE Transactions on Communications 32(4), 396 – 402 (Apr 1984)
4. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley and Sons, Inc., New York (2001)
5. Duin, R., Juszczak, P., Paclík, P., Pękalska, E., de Ridder, D., Tax, D., Verzakov, S.: PR-Tools4.1, a Matlab toolbox for pattern recognition
6. Efron, B.: Bootstrap methods: another look at the jacknife. Annals Statistics 7, 1 – 26 (1979)
7. 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)
8. Kukushkina, O.V., Polikarpov, A.A., Khmelev, D.V.: Using literal and grammatical statistics for authorship attribution. Problems of Information Transmission 37(2), 172–184 (2001)
9. 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)
10. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Transactions on Information Theory 22, 75–81 (1976)
11. Li, M., Chen, X., Li, X., Ma, B., Vitányi, P.M.B.: The similarity metric. IEEE Transactions on Information Theory 50(12), 3250–3264 (2004)
12. Li, M., Vitányi, P.M.B.: An Introduction to Kolmogorov Complexity and its Applications. Springer-Verlag, New York (1997)
13. Mahoney, M.: Large text compression benchmark, http://www.mattmahoney.net/text/text.html
14. Pękalska, E., Skurichina, M., Duin, R.: Combining fisher linear discriminants for dissimilarity representations. In: Proceedings of the First International Workshop on Multiple Classifier Systems. vol. 1857, pp. 117–126. Springer-Verlag (2000)
15. van Rijsbergen, C.: Information Retrieval. Butterworth (1979)
16. 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)

17. Shkarin, D.: PPM: one step to practicality. In: Proceedings of the Data Compression Conference. vol. DDC '02, p. 202. IEEE Computer Society (2002)
18. Telles, G., Minghim, R., Paulovich, F.: Normalized compression distance for visual analysis of document collections. Computers and Graphics 31(3), 327 – 337 (2007)
19. Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval 1, 69–90 (1999)