Author Identification using Compression Models

D. Pavelec[†], L. S. Oliveira[♮], E. Justino[†], F. D. Nobre Neto[‡], and L. V. Batista[‡]

[†]Pontifícia Universidade Católica do Paraná, Curitiba, Brazil
[♮]Universidade Federal do Paraná, Curitiba, Brazil
[‡]Universidade Federal da Paraíba, João Pessoa, Brazil

Abstract

In this paper we discuss the use of compression algorithms for author identification. We present the basic background about compression algorithms and introduce the Prediction by Partial Matching algorithm, which has been used in our experiments. To better compare the results produced by the PPM algorithm, we present some experiments using stylometric features used very often by forensic examiners. In this case the authors are modeled using Support Vector Machines. Comprehensive experiments performed on a database composed of 20 different authors show that the PPM algorithm is an interesting alternative for author identification, since all the process of feature definition, extraction, and selection can be avoided.

1 Introduction

The literature shows a long history of linguistic and stylistic investigation into author identification [8], [7] but the work published by Svartvik [13] marked the birth of term forensic linguistics, i.e., the linguistic investigation of authorship for forensic purposes. In it, he analyzed four statements that Timothy Evans, executed in 1950 for the murder of his wife and baby daughter, was alleged to have made following his arrest. Using both qualitative and quantitative methods Svartvik demonstrated considerable stylistic discrepancies between the statements, thus raising serious questions about their authorship. It was later discovered that both victims had actually been murdered by Evan's landlord, John Christie [2].

Since then, there has been a impressive growth in the volume with which lawyers and courts have called upon the expertise of linguists in cases of disputed authorship. Hence, practical applications for author identification have grown in several different areas such as, criminal law (identifying writers of ransom notes and harassing letters), civil law (copyright and estate disputes), and computer security (mining email content).

Author identification is the task of identifying the author of a given text, therefore, it can be formulated as a typical classification problem, which depends on discriminant features to represent the style of an author. In this context, stylometric features which include various measures of vocabulary richness and lexical repetition based on word frequency distributions, play an important role. As observed by Madigan et al [6], most of these measures, however, are strongly dependent on the length of the text being studied, hence, are difficult to apply reliably. Many other types of features have been investigated, including word class frequencies, syntactic analysis, word collocations, grammatical errors, number of words, sentences, clauses, and paragraph lengths [3, 5, 1].

In this work we discuss an alternative strategy which avoids defining features explicitly while describing the classes as a whole. The feature extraction is performed by a modern lossless data compression algorithm, the PPM Prediction by Partial Matching (PPM) algorithm [9], which is considered one of the best general-purpose compressing algorithm. Despite demanding much more computer resources than dictionary-based techniques, PPM typically yields substantially improved compression ratios. With modern digital technology, memory usage and processing time of PPM is acceptable.

Those that speak in defense of this strategy argue that it yields an overall judgement on the document as a whole, rather than discarding information by pre-selecting features and it avoids the messy and rather artificial problem of defining word boundaries [4]. The contrary argument, on the other hand, relies on the fact that features are defined in a black box of which the inner workings are unclear because it does not follow well established forensic protocols. In this paper we show through experimentation that compression algorithm like PPM is an interesting alterative for author identification producing results comparable to stylometry-based classifiers [10]. This is demonstrated by experiments on a database composed of 600 short articles written in Por-

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.208

IEEE computer society

tuguese by 20 different writers.

This paper is organized as follows. Section 2 introduces the basics about compression algorithms and how they can be used for classification. Section 3 describes the database used in the experiments reported in Section 4. Finally, Section 5 concludes this work.

2 Compression Algorithm

2.1 Background

Let S be a stationary discrete information source that generates messages over a finite alphabet $A = \{a_1, a_2, \ldots, a_M\}$. The source chooses successive symbols from A according to some probability distribution that depends, in general, on preceding selected symbols. A generic message will be modeled as a stationary stochastic process $x = \ldots x_{-2}, x_{-1}, x_0, x_1, x_2$, with $x_i \in A$. Let $x^n = \{x_1, x_2, \ldots, x_n\}$ represent a message of length n. Since $|A| = M$, the source can generate M^n different messages of length n. Let $x_i^n, i = \{1, 2, \ldots, M^n\}$ denote the i^{th} of these messages, according to some sorting order, and assume that the source follows a probability distribution P, so that message x_i^n is produced with probability $P(x_i^n)$.

Let

$$G_n(P) = -\frac{1}{n} \sum_{i=1}^{M^n} P(x_i^n) \log_2 P(x_i^n) \frac{\text{bits}}{\text{symbol}} \qquad (1)$$

It can be shown that $G_n(P)$ decreases monotonically with n [12] and the entropy of the source is given by Equation 2

$$H(P) = \lim_{n \to \infty} G_n(P) \frac{\text{bits}}{\text{symbol}} \qquad (2)$$

An alternative formulation for $H(P)$ uses conditional probabilities. Let $P(x_i^{n-1}, a_j)$ be the probability of the sequence $x_i^n = (x_i^{n-1}, a_j)$, i.e., the probability of x_i^{n-1} concatenated with symbol $x_n = a_j$, and let $P(a_j|x_i^{n-1}) = P(x_i^{n-1}, a_j)P(x_i^{n-1})$ be the probability of symbol $x_n = a_j$ given x_i^{n-1}. The entropy of the n^{th} order approximation to $H(P)$ is given by Equation 3.

$$F_n(P) = -\sum_{i=1}^{M^n} \sum_{j=1}^{M} P(x_i^{n-1}, a_j) \log_2 P(a_j|x_i^{n-1}) \frac{\text{bits}}{\text{symbol}} \qquad (3)$$

$F_n(P)$ decreases monotonically with n [12] and the entropy of the source is given by Equation 4

$$H(P) = \lim_{n \to \infty} \log_2 P(a_j|x_i^{n-1}) \frac{\text{bits}}{\text{symbol}} \qquad (4)$$

Equation 4 involves the estimation of probabilities conditioned on an infinite sequence of previous symbols. In practice, finite memory is assumed, and the sources are modeled by an order-$(n-1)$ Markov process, so that P$(a_j| \ldots x_{-1}, x_0, x_1, \ldots x_{n-1} = P(a_j|x_1, \ldots x_{n-1})$.). In this case, $H(P) = F_n(P)$.

The concept of entropy as a measure of information is central to Information Theory [12], and data compression provides an intuitive perspective to the concept. Define the coding rate of a coding scheme as the average number of bits per symbol the scheme uses to encode the output of a source. A lossless compressor is a uniquely decodable coding scheme whose goal is to achieve a coding rate as small as possible. The coding rate of any uniquely decodable coding scheme is always greater than or equal to the source entropy. Optimum coding schemes have a coding rate equal to the theoretical lower bound $H(P)$, thus achieving maximum compression.

For order-$(n-1)$ Markov processes, optimum encoding is reached if and only if a symbol $x_n = a_j$ occurring after x_i^{n-1} is coded with $-\log_2 P(a_j|x_i^{n-1})$ bits [15, 22]. However, it may be impossible to accurately estimate the conditional distribution $P(.|x_i^{n-1})$ for large values of n, due to the exponential growth of the number of different contexts, which brings well-known problems, such as context dilution.

2.2 The PPM Algorithm

Even though the source model P is generally unknown, it is possible to construct a coding scheme based upon some implicit or explicit probabilistic model Q that approximates P. The better Q approximates P, the smaller the coding rate achieved by the coding scheme.

In order to achieve low coding rates, modern lossless compressors rely on the construction of sophisticated models that closely follows the true source model. Statistical compressors, such as PPM, encode messages according to an estimated statistical model for the source.

For stationary sources, PPM algorithm learns a progressively better model during encoding. Many experimental results show that the superiority of the compression performance of PPM, in comparison with other asymptotically optimum compressors, results mainly from its ability to construct a good model for the source in very early stages of the compression process. In other words, PPM constructs ("learns") an efficient model for the message to be compressed faster than its competitors.

The PPM algorithm is based on context modeling and prediction. The PPM starts with a "complete ignorance model" (assuming independent equiprobable variables) and adaptively updates this model as the symbols in the uncompressed stream are coded. Based on the whole sequence of symbols already coded, the model estimates probability distributions for the next symbols, conditioned on a sequence of k previous symbols in the stream. The number of symbols in the context, k, determines the order of the model.

The next symbol, x, is coded by arithmetic coding, with the probability of x conditioned on its context. If x has not previously occurred in that specific context, no estimate for its probability is available. In this case, a special symbol ("escape") is coded, and PPM-C tries do code x in a reduced context, with $k - 1$ antecedent symbols. This process is repeated until a match is found, or the symbol is coded using the independency-equiprobability model.

Experimentation shows that the compression performance of PPM increases as the maximum context size increases up to a certain point, after which performance starts do degrade. This behavior can be explained by the phenomenon of context-dilution and the increased emission of escape symbols. The context size in which compression performance is optimal depends on the message to be compressed, but typical values usually are in the interval 4 to 6.

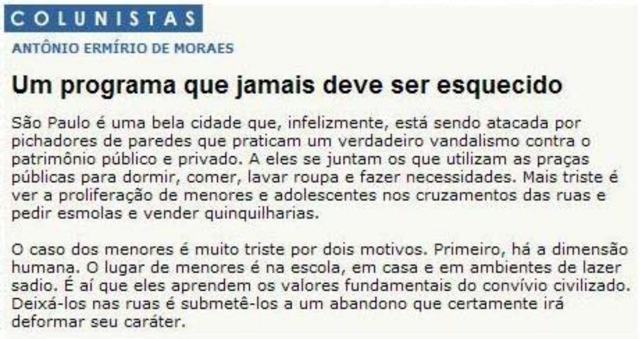
3 Database

To build the database we have collected articles available in the Internet from 20 different people with profiles in Economics (7), Politics (4), Sports (2), Literature (3), Miscellaneous (3), Gossip (1), and Wine (1). Our sources were two different Brazilian newspapers, *Gazeta do Povo* and *Tribuna do Paraná*. We have chosen 30 short articles from each writer. The articles usually deal with polemic subjects and express the author's personal opinion. In average, the articles have 600 tokens and 350 Hapax. The option for short articles was made because in real life forensic experts can count only on short pieces of texts to identify a given writer. Another aspect worth of remark is that this kind of articles can go through some revision process, which can remove some personal characteristics of the texts. Figure 1 depicts an example of the article of our database.

All texts were preprocessed to eliminate numbers, punctuation and diacritics (cedilla, acute accents, etc). Spaces and end-of-line characters are not considered. All hyphenated words are considered as two words. In the example, the sentence "*eu vou dar-te um pula-pula e também dar-te-ei um beijo, meu amor*!" has 16 tokens and 12 Hapax. Punctuation, special characters, and numbers are not considered as tokens.

4 Experiments

In this section we report the experiments we have performed using PPM algorithm. To give a better perspective about the results achieved by this strategy, we also report some results using an SVM classifier trained with stylometric features, which is described in details in Ref. [10].



COLUNISTAS

ANTÔNIO ERMÍRIO DE MORAES

Um programa que jamais deve ser esquecido

São Paulo é uma bela cidade que, infelizmente, está sendo atacada por pichadores de paredes que praticam um verdadeiro vandalismo contra o patrimônio público e privado. A eles se juntam os que utilizam as praças públicas para dormir, comer, lavar roupa e fazer necessidades. Mais triste é ver a proliferação de menores e adolescentes nos cruzamentos das ruas e pedir esmolas e vender quinquilharias.

O caso dos menores é muito triste por dois motivos. Primeiro, há a dimensão humana. O lugar de menores é na escola, em casa e em ambientes de lazer sadio. É aí que eles aprendem os valores fundamentais do convívio civilizado. Deixá-los nas ruas é submetê-los a um abandono que certamente irá deformar seu caráter.

Figure 1. An example of an article used in this work.

Due to the rather small size of the corpus, cross-validation was adopted for computing classification rates. The articles of each author were randomly grouped in three sets (10 samples in each set). In the first cross-validation round, the first set of each author was used for training, and the remaining sets were used for classification. A similar procedure was done in the second and third rounds, with the second and third sets of each author, respectively, selected for training.

In the learning stage, the number N of classes (in this case $N = 20$ writers) is defined, and a training set T_i of text samples with fixed size known to belong to class $C_i, i = \{1, 2, \ldots, N\}$, is selected. In the compression strategy, the feature extraction is done intrinsically by the PPM algorithm. It sequentially compresses the samples in T_i , and the resulting model M_i is kept as a model for the texts in $C_i, i = \{1, 2, \ldots N\}$.

In the classification stage PPM operates in static mode, i.e., the models generated in the training stage are used but not updated during the encoding process. Classification is done as follows: A text sample x from an unknown writer is coded by the PPM algorithm with static model M_i, and the corresponding coding rate $r_i, i = \{1, 2, \ldots N\}$ is registered. Then, the sample x is assigned to C_i if $r_i < r_j, j = \{1, 2, \ldots, N\}, j \neq i$. The rationale is that if x is a sample from class C_i, the model M_i probably best describes its structure, thus yielding the smallest coding rates. The average performance of this strategy on the three different partitions used as testing set (20 articles × 20 authors), was 84.3%. Table 1 shows the confusion matrix produced by this classification scheme.

Table 1 shows that some authors have very strong features for the compression strategy, even when writing about similar subjects, e.g., authors K, O, and Q. Others, like E

Table 1. Confusion matrix produced by the PPM strategy.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A (Literature)	**0.20**	0.13		0.20									0.03		0.03		0.03	0.38		
B (Politics)		**0.87**			0.07		0.03													0.03
C (Economics)			**0.94**					0.03									0.03			
D (Literature)	0.03	0.07		**0.84**											0.03			0.03		
E (Reviewer)					**1.00**															
F (Politics)						**0.97**	0.03													
G (Politics)		0.03					**0.94**			0.03										
H (Economics)		0.03	0.03		0.07			**0.81**							0.03		0.03			
I (Sports)		0.03							**0.97**											
J (Economics)		0.03					0.03			**0.84**										0.10
K (Economics)											**1.00**									
L (Misc)	0.10	0.03										**0.67**			0.03			0.03		0.14
M (Misc)		0.14			0.03				0.14				**0.57**		0.06	0.06				
N (Sports)									0.03					**0.97**						
O (Literature)															**1.00**					
P (Gossip)																**1.00**				
Q (Economics)																	**1.00**			
R (Politics)				0.03							0.03							**0.91**		0.03
S (Economics)											0.30								**0.70**	
T (Economics)		0.03		0.03						0.07	0.07							0.07		**0.74**

and P, write about very specific subjects using a particular vocabulary and for this reason are not misclassified. On the other hand, this strategy achieve a very poor performance for other writers. The most critical case is author A, which writes about literature and had most of their articles confused with author O, which also is a literature critic. Other problems are related to those writes classified as "Misc", which are generalist and write just about everything. Their texts are confused with all sort of writers.

Regarding the experiments based on stylometry, the same formalism has been applied. However, a machine learning algorithm, the well-known Support Vector Machine (SVM), was used to model the N classes (authors). When using SVMs, there are two basic approaches to solve an N-class problems: pairwise and one-against-others. In this work both strategies have been tried out but the former produced better results. A Gaussian kernel was employed and its parameters (C and γ) were defined through a grid search. A modified version of the SVM [11], which is able to produce a estimation the the posterior probability $P(class|input)$ was considered in this work.

The feature vector used to train the SVMs is composed of 171 components, which is the number of occurrences of 77 conjunctions and 94 adverbs found in the text. Conjunctions and adverbs were assessed independently, but with no further improvements. In the classification stage, a text sample x from an unknown author is assigned to C_i that maximizes the posterior probability, i.e., $C_i = \max P(C_i|x)$. The average performance of this strategy on the testing set was 83.2%.

Table 2 reports the average performance of both strategies based on three different partitions of the database. Both achieve similar performances but PPM features a higher standard deviation.

Table 2. Comparison between PPM (compression) and SVM (stylometry) on three different testing partitions. Standard deviation in parenthesis.

Round	PPM (%)	SVM (%)
1	84.0 (22.1)	83.0 (15.2)
2	83.0 (23.4)	84.0 (14.8)
3	86.0 (22.1)	82.9 (15.6)
Average	84.3 (22.5)	83.3 (15.2)

Due to space constraints we are not able to show the confusion matrix produced by the stylometry-based classifier. However, observing the confusion matrices we can notice that, in spite of the similar overall performance, the methods make different confusions. One example is writer A, where PPM performs very poorly (about 20%) and the stylometry-based classifier identify almost all samples of that writer correctly. The opposite situation happens for author O, which is identified correctly 100% by PPM and only 60% by the stylometry-based classifier. Besides, even for similar performances, author D for example, the confusions are different. All that is a good indicative that both methods produce complementary results and can be further combined to build a more reliable identification system.

From the experimental point of view, both strategies have advantages and drawbacks. Compared to the traditional classification scheme used by the stylometric-based classifier, the PPM has some advantages, such as i) no definition of features, ii) no feature extraction and iii) no traditional learning phase. It is worth of remark that in spite of the apparent black box concept, compression algorithms like PPM are based on robust probabilistic frameworks. However, if the size of the text or the number of samples per writer cannot be augmented, the performance of the PPM strategy cannot be further improved. In other words, this is as good as it gets.

On the other hand, the traditional way of defining and extracting features to train a machine learning model gives us more perspective for improvements, since we always can explore new features, select the relevant and uncorrelated ones through feature selection, and try new classification algorithms.

5 Conclusion

In this work we have discussed compression algorithms for author identification, where the feature extraction is done in an implicity fashion where each writer of the database is modeled by compressing some samples of his writings. Then, during recognition those models are used to compress a given questioned samples which is assigned to the class that produces the lowest compression rate. The algorithm used in our experiments was the PPM compression algorithm.

To give a better insight about the usefulness of the PPM for author identification, we have reproduced some results produced by a traditional pattern recognition system, which involves steps such as definition of features, feature extraction, classifier training, etc. In this case, the system takes into account stylometric features (conjunctions and adverbs of the Portuguese language) to train an SVM classifier. Results using the same testing protocol show that both strategies produce very similar results, but making different confusions. This shows that both strategies are complementary to each other and can be combined to build a more reliable identification system. Besides, we believe that the PPM is an useful tool that can be used as parameter when designing new features for author identification. It is fair to expect that a discriminant feature set would at least achieve the same level of performance than PPM.

As future works, we plan to increase the database with more authors and also longer articles, which will enable us to assess the impacts of bigger databases on PPM. Moreover, different stylometric features used by forensic experts will be tried out, as well as strategies to combine these two different paradigms will be investigated.

Acknowledges

This research has been supported by The National Council for Scientific and Technological Development (CNPq) grant 471496/2007-3. Francisco D. Nobre Neto was supported by scholarship from the Tutorial Education Programme (PET) from the Higher Education Secretariat of the Ministry of Education.

References

[1] S. Argamon, M. Saric, and S. S. Stein. Style mining of electronic messages for multiple author discrimination. In *ACM Conference on Knowledge Discovery and Data Mining*, 2003.

[2] M. Coulthard. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4):431–447, 2004.

[3] R. S. Forsyth and D. I. Holmes. Feature finding for text classfication. *Literary and Linguistic Computing*, 11(4):163–174, 1996.

[4] E. Frank, C. Chui, and I. H. Witten. Text categorization using compression models. In *Data Compression Conference*, 2000.

[5] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[6] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye. Author identification on the large scale. In *Joint Annual Meeting of the Interface and the Classification Society of North America (CSNA)*, 2005.

[7] C. Mascol. Curves of pauline and pseudo-pauline style i. *Unitarian Review*, 30:453–460, 1888.

[8] T. Mendenhall. The characteristic curves of composition. *Science*, 214:237–249, 1887.

[9] A. Moffat. Implementing the PPM data compression scheme. *IEEE Transactions on Communications*, 38(11):1917–1921, 1990.

[10] D. Pavelec, E. Justino, L. V. Batista, and L. S. Oliveira. Author identification using writerdependent and writerindependent strategies. In *ACM Symposium on Applied Computing*, pages 414–418, 2008.

[11] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. S. et al, editor, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[12] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. Journal*, 27:379–423, 1948.

[13] J. Svartvik. The evans statements: A case for forensic linguistics. *Acta Universitatis Gothoburgensis*, 1968.