Information Theoretic Text Classification Methods Evaluation

David Pereira Coutinho[1] and Mário A. T. Figueiredo[2]

[1] Depart. de Engenharia de Electrónica e Telecomunicações e de Computadores
Instituto Superior de Engenharia de Lisboa, 1959-007 Lisboa, Portugal

[2] Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal

Abstract. Most approaches to text classification rely on some measure of (dis)similarity between sequences of symbols. Information theoretic measures have the advantage of making very few assumptions on the models which are considered to have generated the sequences, and have been the focus of recent interest. This paper compares the use of the *Ziv-Merhav method* (ZMM) and the *Cai-Kulkarni-Verdú method* (CKVM) for the estimation of relative entropy (or Kullback-Leibler divergence) from sequences of symbols when used as a tool for text classification. We describe briefly our implementation of the ZMM based on a modified version of the *Lempel-Ziv algorithm* (LZ77) and also the CKVM implementation which is based in the *Burrows-Wheeler block sorting transform* (BWT). Assessing the accuracy of both the ZMM and CKVM on synthetic Markov sequences shows that CKVM yields better estimates of the Kullback-Leibler divergence. Finally, we apply both methods in a text classification problem (more specifically, authorship attribution) but surprisingly CKVM permforms poorly while ZMM outperforms a previously proposed (also information theoretic) method.

1 Introduction

Defining a similarity measure between two finite sequences, without explicitly modelling their statistical behavior, is a fundamental problem with many important applications in areas such as information retrieval or text classification.

Approaches to this problem include: various types of edit (or Levenshtein) distances between pairs of sequences (*i.e.*, the minimal number of edit operations, chosen from a fixed set, required to transform one sequence into the other; see, *e.g.*, [1], for a review); "universal" distances (*i.e.* independent of a hypothetical source model) such as the *information distance* [2]; methods based on universal (in the Lempel-Ziv sense) compression algorithms [3] [4] and on the Burrows-Wheeler block sorting transform [5].

[2] This work was partially supported by Fundao para a Cincia e Tecnologia (FCT), under grant PTDC/EEA-TEL/72572/2006.

Pereira Coutinho D. and A. T. Figueiredo M. (2008).
Information Theoretic Text Classification Methods Evaluation.
In *Proceedings of the 8th International Workshop on Pattern Recognition in Information Systems*, pages 77-85

In this paper, we consider using the methods proposed by Ziv and Merhav (ZM) [3] and by Cai, Kulkarni and Verdú (CKV) [5] for the estimation of relative entropy, or Kullback-Leibler (KL) divergence, from pairs of sequences of symbols, as a tool for text classification. In particular, to handle the text authorship attribution problem, Benedetto, Caglioti and Loreto (BCL) [4] introduced a "distance" function based on an estimator of the relative entropy obtained by using the *gzip* compressor [6] and file concatenation. This work follows the same idea of estimating a dissimilarity using data compression techniques, but using both the ZM method [7] and the CKV method [5], with the main purpose of comparing these two KL divergence estimators in this context.

We describe briefly our implementation of the ZM method based on a modified version of the Lempel-Ziv algorithm (LZ77) and also the CKV method implementation which is based in the Burrows-Wheeler block sorting transform (BWT) [8]. We assess the accuracy of both the ZM and CKV estimators on synthetic Markov sequences, showing that, for these sources, CKV yields better estimates of the KL divergence. Finally, we apply both ZM and CKV methods to an authorship attribution problem using a text corpus similar to the one used in [4]. Results shows that CKV method permforms poorly while ZM method outperforms the technique introduced in [4].

The outline of the paper is has follows. In Section 2 we recall the fundamental tools used in this approach: the concept of relative entropy and the relationship between entropy and Lempel-Ziv coding. In Section 3 we describe briefly the BCL, ZM and CKV methods. Section 4 presents the experimental results, while Section 5 concludes the paper.

2 Data Compression and Similarity Measures

2.1 Kullback-Leibler Divergence and Optimal Coding

Consider two memoryless sources $\mathcal{A}$ and $\mathcal{B}$ producing sequences of binary symbols. Source $\mathcal{A}$ emits a 0 with probability p (thus a 1 with probability $1-p$) while $\mathcal{B}$ emits a 0 with probability q. According to Shannon [9], there are compression algorithms that applied to a sequence emitted by $\mathcal{A}$ will be asymptotically able to encode the sequence with an average number bits per character equal to the source entropy $H(\mathcal{A})$, *i.e.*, coding, on average, every character with

$$H(\mathcal{A}) = -p\,\log_2 p - (1-p)\,\log_2(1-p) \quad \text{bits.} \tag{1}$$

An optimal code for $\mathcal{B}$ will not be optimal for $\mathcal{A}$ (unless, of course, $p=q$). The average number of extra bits per character which are wasted when we encode sequences emitted by $\mathcal{A}$ using an optimal code for $\mathcal{B}$ is given by the relative entropy (KL divergence) between $\mathcal{A}$ and $\mathcal{B}$ (see, *e.g.*, [9]), that is

$$D(\mathcal{A}||\mathcal{B}) = p\log_2\frac{p}{q} + (1-p)\log_2\frac{1-p}{1-q}. \tag{2}$$

The observation in the previous paragraph points to the following possible way of estimating the KL divergence between two sources: design an optimal code for source $\mathcal{B}$ and measure the average code-length obtained when this code is used to encode

sequences from source $\mathcal{A}$. The difference between this average code-length and the entropy of $\mathcal{A}$ provides an estimate of $D(\mathcal{A}||\mathcal{B})$. Of course, the entropy of $\mathcal{A}$ itself can be estimated by measuring the average code-length of an optimal code for this source. This is the basic rationale underlying the approaches proposed in [4], [3], and [5]. However, to use this idea for general sources (not simply for the memoryless ones that we have considered up to now for simplicity), without having to explicitly estimate models for each of them, we need to use some form of universal coding. A universal coding technique (such as Lempel-Ziv [10] or BWT-based coding [11]) is one that is able to asymptotically achieve the entropy rate lower bound without prior knowledge of a model of the source (which, of course, does not have to be memoryless) [9].

2.2 Relationship between Entropy and Lempel-Ziv Coding

Assume a random sequence $\mathbf{x} = (x_1, x_2, ..., x_n)$ was produced by an unknown order-n stationary Markovian source, with a finite alphabet. Consider the goal estimating the *n*th-order entropy, or equivalently the logarithm of the joint probability function $-(1/n)\log_2 p(x_1, x_2, ..., x_n)$ (from which the entropy could be obtained). A direct approach to this goal is computationally prohibitive for large n, or even impossible if n is unknown. However, an alternative route can be taken using the following fact (see [9], [12]): the Lempel-Ziv (LZ) code length for **x**, divided by n, is a computationally efficient and reliable estimate of the entropy, and hence also of $-(1/n)\log_2 p(x_1, x_2, ..., x_n)$. More formally, let $c(\mathbf{x})$ denote the number of phrases in **x** resulting from the LZ incremental parsing of **x** into distinct phrases, such that each phrase is the shortest sequence which is not a previously parsed phrase. Then, the LZ code length for **x** is approximately

$$c(\mathbf{x})\log_2 c(\mathbf{x}) \tag{3}$$

and it can be shown that this quantity converges (with n) almost surely to $-(1/n)$ $\log_2 p(x_1, x_2, ..., x_n)$, as $n \to \infty$ [3]. This fact suggests using the output of an LZ encoder to estimate the entropy of an unknown source without explicitly estimating its model parameters.

3 Information Theoretic Methods

3.1 The Method of Benedetto, Caglioti and Loreto

Benedetto *et al* [4] have proposed a particular way of using LZ coding to estimate the KL divergence between two sources (in fact, sequences) $\mathcal{A}$ and $\mathcal{B}$. They have used the proposed method for context recognition and for classification of sequences. In this subsection, we briefly review their method.

Let $|X|$ denote the length in bits of the uncompressed sequence X, let L_X denote the length in bits obtained after compressing sequence X (in particular, [4] uses *gzip*, which is an LZ-based compression algorithm [6]), and let $X + Y$ stand for the concatenation of sequences X and Y (with Y after X). Let A and B be "long" sequences from sources

$\mathcal{A}$ and $\mathcal{B}$, respectively, and let b be a "short" sequence from source $\mathcal{B}$. As proposed by Benedetto *et al*, the relative entropy $D(\mathcal{A}||\mathcal{B})$ (per symbol) can be estimated by

$$\widehat{D}(\mathcal{A}||\mathcal{B}) = (\Delta_{Ab} - \Delta_{Bb})/|b|, \tag{4}$$

where $\Delta_{Ab} = L_{A+b} - L_A$ and $\Delta_{Bb} = L_{B+b} - L_B$. Notice that $\Delta_{Ab}/|b|$ can be seen as the code length (per symbol) obtained when coding a sequence from $\mathcal{B}$ (sequence b) using a code optimized for $\mathcal{A}$, while $\Delta_{Bb}/|b|$ can be interpreted as an estimate of the entropy of the source $\mathcal{B}$.

To handle the text authorship attribution problem, Benedetto, Caglioti and Loreto (BCL) [4] defined a simplified "distance" function $d(A, B)$ between sequences,

$$d(A, B) = \Delta_{AB} = L_{A+B} - L_A, \tag{5}$$

which we will refer to as the BCL divergence. As mention before, Δ_{AB} is a measure of the description length of B when the coding is optimized to A, obtained by subtracting the description length of A from the description length of $A+B$. Hence, it can be stated that $d(A, B'') < d(A, B')$ means that B'' is more similar to A than B'. Notice that the BCL divergence is not symmetric.

More recently, Puglisi *et al* [13] studied in detail what happens when a compression algorithm, such as LZ77 [10], tries to optimize its features at the interface between two different sequences A and B, while compressing the sequence $A + B$. After having compressed sequence A, the algorithm starts compressing sequence B using the dictionary that it has learned from A. After a while, however, the dictionary starts to become adapted to sequence B, and when we are well into sequence B the dictionary will tend to depend only on the specific features of B. That is, if B is long enough, the algorithm learns to optimally compress sequence B. This is not a problem when the sequence B is sufficiently short for the the dictionary not to become completely adapted to B, but is a serious problem arises for a long sequence B. The Ziv-Merhav method, described next, does not suffer from this problem, this being what motivated us to consider it for sequence classification problems [7].

3.2 Ziv-Merhav Empirical Divergence

The method proposed by Ziv and Merhav [3] for measuring relative entropy is also based on two Lempel-Ziv-type parsing algorithms:

- The incremental LZ parsing algorithm [12], which is a self parsing procedure of a sequence into $c(\mathbf{z})$ distinct phrases such that each phrase is the shortest sequence that is not a previously parsed phrase. For example, let $n = 11$ and $\mathbf{z} = (01111000110)$, then the self incremental parsing yields $(0, 1, 11, 10, 00, 110)$, namely, $c(\mathbf{z}) = 6$.
- A variation of the LZ parsing algorithm described in [3], which is a sequential parsing of a sequence $\mathbf{z}$ with respect to another sequence $\mathbf{x}$ (cross parsing). Let $c(\mathbf{z}|\mathbf{x})$ denote the number of phrases in $\mathbf{z}$ with respect to $\mathbf{x}$. For example, let $\mathbf{z}$ as before and $\mathbf{x} = (10010100110)$; then, parsing $\mathbf{z}$ with respect to $\mathbf{x}$ yields $(011, 110, 00110)$, that is $c(\mathbf{z}|\mathbf{x}) = 3$.

Ziv and Merhav have proved that for two finite order (of any order) Markovian sequences of length n the quantity

$$\Delta(\mathbf{z}||\mathbf{x}) = \frac{1}{n}\,[\,c(\mathbf{z}|\mathbf{x})\,\log_2 n - c(\mathbf{z})\,\log_2 c(\mathbf{z})\,] \tag{6}$$

converges, as $n \to \infty$, to the relative entropy between the two sources that emitted the two sequences **z** and **x**. Roughly speaking, we can observe (see (3)) that $c(\mathbf{z})\,\log_2 c(\mathbf{z})$ is the measure of the complexity of the sequence **z** obtained by self-parsing, thus providing an estimate of its entropy, while $(1/n)\,c(\mathbf{z}|\mathbf{x})\,\log_2 n$ can be seen as an estimate of the code-length obtained when coding **z** using a model for **x**. From now on we will refer to $\Delta(\mathbf{z}||\mathbf{x})$ as the ZM divergence.

Our implementation of the ZM divergence [7] uses the LZ78 algorithm to make the self parsing procedure. To perform the cross parsing, we designed a modified LZ77-based algorithm where the dictionary is static and only the lookahead buffer slides over the input sequence, as shown in Figure 1.

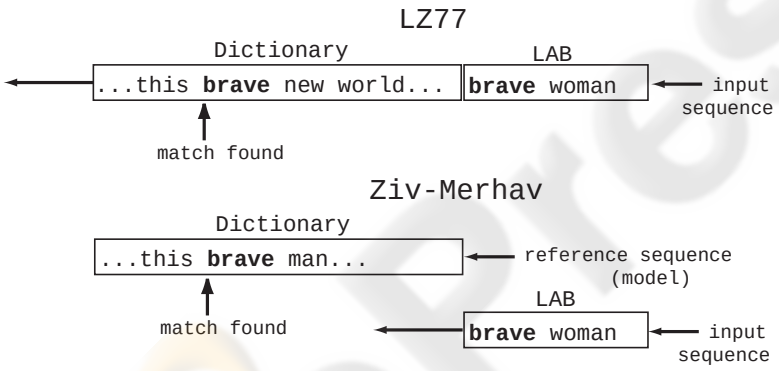

Fig. 1. The original LZ77 algorithm uses a sliding window over the input sequence to get the dictionary updated, whereas in the Ziv-Merhav cross parsing procedure the dictionary is static and only the *lookahead buffer* (LAB) slides over the input sequence.

Two important parameters of the algorithm are the dictionary size and the maximum length of a matching sequence found in the LAB; both influence the parsing results and determine the compressor efficiency [6]. The experiments reported in the Experiments section were performed using a 65536 byte dictionary and a 256 byte long LAB.

3.3 The BWT-based Method

The divergence estimator proposed by Cai, Kulkarni and Verdú applies the Burrows-Wheeler transform (BWT) to the concatenation of the two sequences for which the estimation divergence is wanted.

The BWT is a reversible block-sorting algorithm [8]. It operates on a sequence of symbols, produces all cyclic shifts of the original sequence, sorts them lexicographically, and outputs the last column of the sorted table. For finite-memory sources, performing the BWT on a reversed data sequence groups together symbols in the same state. Using the BWT followed by segmentation is the basic idea behind the entropy estimation in [14]. This idea was extended to divergence estimation [5] introducing the joint BWT of two sequences as illustrated in Figure 2.

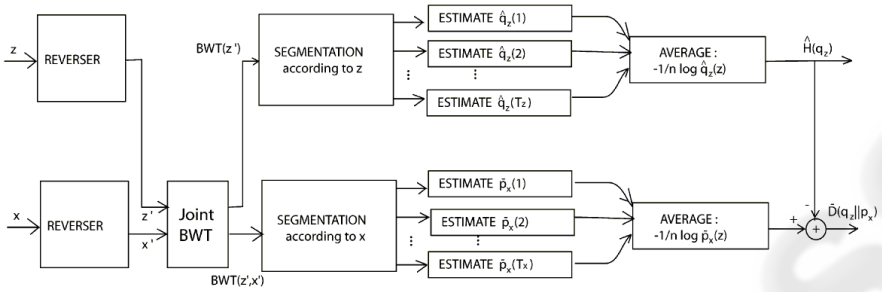

Fig. 2. Block diagram of the divergence estimator via the BWT.

4 Experiments

4.1 Synthetic Data: Binary Sources

The purpose of our first experiments is to compare the theoretical values of the KL divergence with the estimates produed by the ZM and the CKV methods, on pairs of binary sequences with 100, 1000 and 10000 symbols. The sequences were randomly generated from simulated sources using both memoryless and order-1 Markov models. For the memoryless sources, the KL divergence is given by expression (2), while for the order-1 sources it is given by

$$D(p||q) = \sum_{x_1,x_2} p(x_1,x_2)\ \log_2 \frac{p(x_2|x_1)}{q(x_2|x_1)}. \tag{7}$$

Results for these experiments using 10000 symbols are shown in Figure 3. Each plot compares the true KL divergence with the ZM and CKV estimates, over a varying range of source symbol probabilities. The results show that, for this type of source, the CKV method provides a more accurate KL divergence estimate than the ZM technique (which may even return negative values when the sequences are very similar).

4.2 Text Classification

Our next step is to compare the performance of the ZM and CKV estimators of the KL divergence with the BCL divergence on the authorship attribution problem. We use the

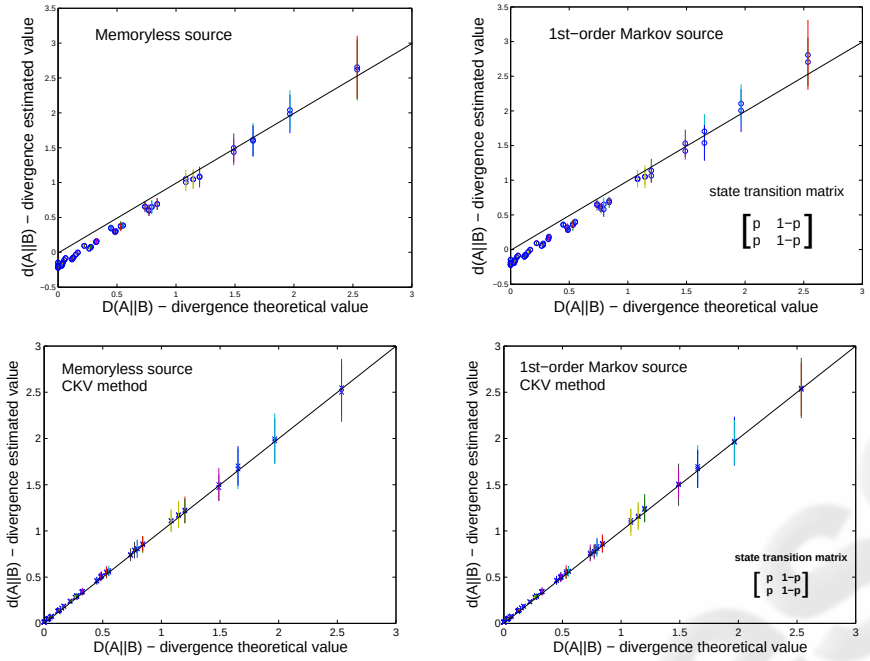

Fig. 3. KL divergence estimates obtained by the ZM and CKV methods, versus the theoretical values. Each circle is the sample mean value and the vertical bars the sample standard deviation values, evaluated over 100 pairs of sequences (of length 10000). For the 1st-order Markov source we use the state transition matrix shown and consider a range of values of $p \in [0, 1]$.

same text corpus that was used by Benedetto *et al* [4]. This corpus contains a set of 86 files of several Italian authors, and can be downloaded from www.liberliber.it. Since we don't know exactly which files were used in [4], we apply also BCL method to this new corpus of Italian authors. In this experiment, each text is classified as belonging to the author of the closest text in the remaining set. In other words, the results reported can be seen as a full *leave-one-out cross-validation* (LOO-CV) performance measure of a nearest-neighbor classifier built using the considered divergence functions.

The results of this experiment, which are presented in Table I, show that the ZM divergence outperforms the other divergences over the very same corpus. Our rate of success using the ZM divergence is 95.4%, while the BCL and the CKV divergence achieves rate of success of 90.7% and 38.4% respectively. Notice that the CKV rate of success will improve to 47.7% if each text is classified as belonging to one of the authors of the two closest texts in the remaining set.

5 Conclusions

We have compared the Cai-Kulkarni-Verdú (CKV) [5] and the Ziv-Merhav (ZM) [3] methods for Kullback-Leibler divergence estimation, and assessed their performance as

Table 1. Classification of Italian authors: for each author, we report the number of texts considered and three values of classification success rate, obtained using the method of Benedetto, Caglioti and Loreto (BCL), the Ziv-Merhav method (ZM) and the method proposed by Cai, Kulkarni and Verdú (CKV).

Author	No. of texts	BCL	ZM	CKV
Alighieri	8	7	7	7
Deledda	15	15	15	0
Fogazzaro	5	3	5	4
Guicciardini	6	6	5	0
Macchiavelli	12	11	11	5
Manzoni	4	4	3	4
Pirandello	11	9	11	3
Salgari	11	11	11	8
Svevo	5	5	5	1
Verga	9	7	9	1
Total	**86**	**78**	**82**	**33**

a tool for text classification. Computational experiments showed that the CKV method yields better estimates of the KL divergence on synthetic Markov sequences. However, when both methods were applied to a text classification problem (specifically, authorship attribution), the CKV method was clearly outperformed by ZM method, which also outperforms the method introduced by Benedetto, Caglioti and Loreto [4]

Future work will include further experimental evaluation on other text classification tasks, as well as the development of more sophisticated text classification algorithms. Namely, we plan to define information-theoretic kernels based on these KL divergence estimators and use them in kernel-based classifiers such as support vector machines [15].

Acknowledgements

The authors would like to thank Haixiao Cai for providing his implementation of the CKV method and for his comments about our work.

References

1. D. Sankoff and J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley, 1983.
2. C. Bennett, P. Gacs, M. Li, P. Vitanyi, and W. Zurek, "Information distance," *IEEE Transactions on Information Theory*, vol. 44, pp. 1407–1423, 1998.
3. J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Transactions on Information Theory*, vol. 39, pp. 1270–1279, 1993.
4. D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters, 88:4*, 2002.
5. H. Cai, S. Kulkarni, and S. Verdu, "Universal divergence estimation for finite-alphabet sources," *IEEE Transactions on Information Theory*, vol. 52, pp. 3456–3475, 2006.

6. M. Nelson and J. Gailly, *The Data Compression Book*. M&T Books, New York, 1995.
7. D. Pereira Coutinho and M. Figueiredo, "Information theoretic text classification using the Ziv-Merhav method," *2nd Iberian Conference on Pattern Recognition and Image Analysis – IbPRIA'2005*, 2005.
8. M. Burrows and D. Wheeler, "A block-sorting lossless data compression algorithm," *Tech. Rep. 124, Digital Systems Research Center*, 1994.
9. T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.
10. J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 337–343, 1977.
11. M. Effros, K. Visweswariah, S. Kulkarni, and S. Verdu, "Universal lossless source coding with the Burrows-Wheeler transform," *IEEE Trans. on Information Theory*, vol. 48, pp. 1061–1081, 2002.
12. J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530–536, 1978.
13. A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani, "Data compression and learning in time sequences analysis," *Physica D*, vol. 180, p. 92, 2003.
14. H. Cai, S. Kulkarni, and S. Verdu, "Universal estimation of entropy via block sorting," *IEEE Transactions on Information Theory*, vol. 50, pp. 1551–1561, 2004.
15. J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Recognition*. Cambridge University Press, 2004.