Information Theoretic Text Classification Using the Ziv-Merhav Method

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
davidpc@isel.pt

[2] Instituto de Telecomunicações
Instituto Superior Técnico
1049-001 Lisboa, Portugal
mtf@lx.it.pt

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 addresses the use of the *Ziv-Merhav method* (ZMM) for the estimation of relative entropy (or Kullback-Leibler divergence) from sequences of symbols as a tool for text classification. We describe an implementation of the ZMM based on a modified version of the Lempel-Ziv algorithm (LZ77). Assessing the accuracy of the ZMM on synthetic Markov sequences shows that it yields good estimates of the Kullback-Leibler divergence. Finally, we apply the method in a text classification problem (more specifically, authorship attribution) outperforming 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].

In this paper, we consider using the method proposed by Ziv and Merhav (ZM) 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

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 355–362, 2005.

Loreto [3] introduced a "distance" function based on an estimator of the relative entropy obtained by using the *gzip* compressor [4] and file concatenation. This work follows the same idea of estimating a dissimilarity using data compression, but using the ZM method [5]. The ZM approach avoids the drawbacks of the method of Benedetto *et al* [3] which have been pointed out by Puglisi *et al* [6], and has desirable theoretical properties of fast convergence.

We describe an implementation of the ZM method based on a modified version of the Lempel-Ziv algorithm. We assess the accuracy of the ZM estimator on synthetic Markov sequences, showing that it yields good estimates of the KL divergence. Finally, we apply the method to an authorship attribution problem using a text corpus similar to the one used in [3]. Our results show that ZM method outperforms the technique introduced in [3].

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, the method proposed by Bennedeto *et al*, and the ZM method. In Section 3 we describe our implementation of the ZM technique based on the LZ77 algorithm. 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 [7, 8], 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.*, [8]), that is

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

This fact suggests the following possible way to estimate the KL divergence between two sources: design an optimal code for source $\mathcal{B}$ and then measure the average number of bits obtained when this code is used to encode sequences from source $\mathcal{A}$. The difference between this average code length and the entropy of $\mathcal{A}$ is an estimate of the KL divergence $D(\mathcal{A}||\mathcal{B})$. The entropy of $\mathcal{A}$ itself can be estimated by measuring the average code length of an adapted optimal code. This is the basic idea that underlies the methods proposed in [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 the Lempel-Ziv algorithm) is one that is asymptotically able to achieve the entropy lower bound without prior knowledge of the source distribution (which, of course, does not have to be memoryless) [8].

2.2 Relationship Between Entropy and Lempel-Ziv Coding

Consider a sequence $\mathbf{x} = (x_1, x_2, ..., x_n)$ emitted by an unknown lth-order stationary Markovian source, defined over a finite alphabet. Suppose that one wishes to estimate the nth-order entropy, or equivalently $-(1/n)\log_2 p(x_1, x_2, ..., x_n)$. A direct approach to this goal is computationally prohibitive for large l, or even impossible if l is unknown. However, an alternative route can be taken using the following fact (see [8], [9]): the Lempel-Ziv (LZ) code length for $\mathbf{x}$, 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 $\mathbf{x}$ resulting from the LZ sequential parsing of $\mathbf{x}$ into distinct phrases, such that each phrase is the shortest sequence which is not a previously parsed phrase. Then, the LZ code length for $\mathbf{x}$ can be approximated by

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

and it can be shown that it converges almost surely to $-(1/n)\log_2 p(x_1, x_2, ..., x_n)$, as $n \to \infty$ [5]. This shows that we can use the output of an LZ encoder to estimate the entropy of an unknown source without explicitly estimating its model parameters.

2.3 The Method of Benedetto, Caglioti and Loreto

Recently, Benedetto *et al* [3] have proposed a particular way of using LZ coding to estimate KL divergence between two sources $\mathcal{A}$ and $\mathcal{B}$. They have used the proposed method for context recognition and classification of sequences.

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, [3] uses *gzip*, which is an LZ-based compression algorithm [4]), 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 b a "small" sequence from source $\mathcal{B}$. As proposed by Benedetto *et al*, the relative entropy $D(\mathcal{A}||\mathcal{B})$ (per character) 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 character) 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) [3] 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* [6] 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 so short that the dictionary does not become completely adapted to B. In this case, one can measure the relative entropy by compressing the sequence $A+B$. The problem arises for long sequences 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.

2.4 Ziv-Merhav Empirical Divergence

The method proposed by Ziv and Merhav [5] for measuring relative entropy is also based on two Lempel-Ziv-type parsing algorithms:

- The incremental LZ parsing algorithm [9], 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 [5], 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

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

converges, as $n \to \infty$, to the relative entropy between the two sources that emitted the two sequences $\mathbf{z}$ and $\mathbf{x}$. 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 $\mathbf{z}$ 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 $\mathbf{z}$ using a model for $\mathbf{x}$. From now on we will refer to $\Delta(\mathbf{z}||\mathbf{x})$ as the ZM divergence.

3 Modified LZ77 Algorithm

We have implemented the ZM divergence using 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. For better understanding, let us briefly recall the LZ77 algorithm and its implementation model.

The LZ77 compression algorithm observes the input sequence through a sliding window buffer as shown in Figure 1. The sliding window buffer consists of a dictionary and a *lookahead buffer* (LAB). The dictionary holds the symbols already analyzed and the LAB the symbols to be analyzed. At each step, the algorithm tries to express the sequence in the LAB as a subsequence in the dictionary using a reference to it and then coding that match. Otherwise, the leftmost symbol in the LAB is coded as a literal. In both situations, the dictionary is updated after each step.

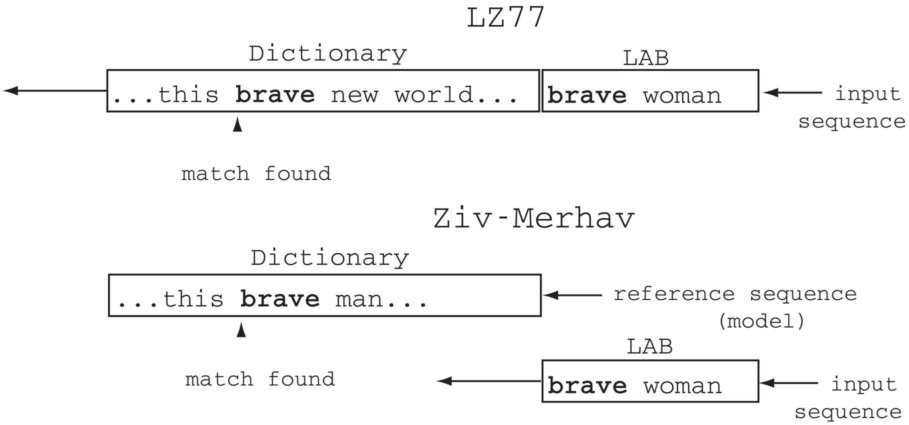

Fig. 1. The original LZ77 algorithm uses a sliding window over the input sequence to get the dictionary updated, whereas in the Ziv-Merhav cross parsing procedure the dictionary is static and only the *lookahead buffer* (LAB) slides over the input sequence.

To implement the cross parsing procedure, we first use the reference sequence (model) to build an LZ77-like dictionary, which will remain static. After that, the input sequence (to be compared) slides through the LAB from right to left as shown in Figure 1. At each step, the procedure is the same as with LZ77, except that the dictionary is not updated.

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 [4]. The experiments reported in the next section were performed using a 65536 byte dictionary and a 256 byte long LAB.

4 Experiments

4.1 Synthetic Data

The purpose of our first experiments was to compare the theoretical values of the KL divergence with the estimates produced by the ZM method, on pairs of binary sequences with 100, 1000 and 10000 symbols. The sequences were randomly generated from simulated sources using 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 are shown in Figure 2. Each experiment compares KL divergence against ZM divergence, over a varying range of source symbol probabilities. The results show that the ZM divergence provides a good KL divergence estimate, regardless its negative values when the sequences are very similar or "close".

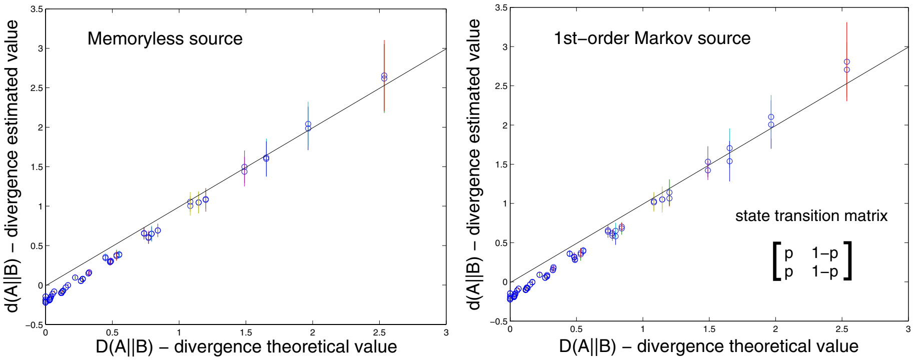

Fig. 2. Theoretical values versus Ziv-Merhav empirical divergence values, between two synthetic binary sequences of 10000 symbols length. Each circle is the sample mean value and the vertical segments are the sample standard deviation values, evaluated over 100 sequence pairs. For the 1st-order Markov source we use the state transition matrix shown and test for all probabilities $p \in [0, 1]$. Results are near to the identity line of no estimation error.

4.2 Text Classification

Our next step was to compare the performance of ZM divergence with the BCL divergence on the authorship attribution problem using a text corpus similar to the one used by Benedetto *et al* [3]. For this purpose, we have used a set of 86 files of the same authors, downloaded from the same site: `www.liberliber.it`. Since we don't know exactly which files were used in [3], we apply both measures 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.

Table 1. Italian Authors Classification - For each author we report the number of texts considered and two measures of classification success, one obtained using the original method proposed by Benedetto, Caglioti and Loreto (BCL) and the other with the Ziv-Merhav method (ZM).

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

The results of this experiment, which are presented in Table I, show that the ZM divergence outperforms the BCL divergence over the very same corpus. Our rate of success using the ZM divergence is 95.4%, while the BCL divergence achieves rate of success of 90.7%.

5 Conclusion

We have presented an implementation of the Ziv-Merhav method for the estimation of relative entropy or Kullback-Leibler divergence from sequences of symbols, which can be used as a tool for text classification. Computational experiments showed that this method yields good estimates of the relative entropy on synthetic Markov sequences. Moreover, this method was applied to a text classification problem (authorship attribution), outperforming a previously proposed approach. Future work will include further experimental evaluation of the Ziv-Merhav method, as well as its use in more sophisticated text classification algorithms such as a kernel-based methods [11].

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. D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters, 88:4*, 2002.
4. M. Nelson and J. Gailly, *The Data Compression Book 2nd edition.* M&T Books, New York, 1995.
5. J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. on Information Theory, pp. 1270–1279*, 1993.
6. A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani, "Data compression and learning in time sequences analysis," *Physica D 180, 92*, 2003.
7. C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal,27: pp. 379-423, pp. 623-656*, 1948.
8. T. Cover and J. Thomas, *Elements of Information Theory.* John Wiley & Sons, Inc, 1991.
9. 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.
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. J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Recognition.* Cambridge University Press, 2004.