UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

CLASSIFICATION OF SEQUENCES USING COMPRESSION-BASED DISSIMILARITY MEASURES

José David Pereira Coutinho Gomes Antão

Supervisor: Doctor Mário Alexandre Teles de Figueiredo

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering

Jury final classification: Pass With Merit

Chairperson: CHAIRMAN OF THE IST SCIENTIFIC BOARD
Members of the Committee:

Doctor MÁRIO ALEXANDRE TELES DE FIGUEIREDO

Doctor ARMANDO JOSÉ FORMOSO DE PINHO

Doctor JORGE DOS SANTOS SALVADOR MARQUES

Doctor LUÍS FILIPE COELHO ANTUNES

Doctor ANA LUÍSA NOBRE FRED

2014

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Classification of Sequences using Compression-Based Dissimilarity Measures

José David Pereira Coutinho Gomes Antão

Supervisor: Doctor Mário Alexandre Teles de Figueiredo

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering

Jury final classification: Pass With Merit

Chairperson: Chairman of the IST Scientific Board

Members of the Committee:

Doctor Mário Alexandre Teles de Figueiredo, Professor Catedrático do Instituto Superior Técnico, da Universidade de Lisboa

Doctor Armando José Formoso de Pinho, Professor Associado (com Agregação) da Universidade de Aveiro

Doctor Jorge dos Santos Salvador Marques, Professor Associado (com Agregação) do Instituto Superior Técnico, da Universidade de Lisboa

Doctor Luís Filipe Coelho Antunes, Professor Associado da Faculdade de Ciências, da Universidade do Porto

Doctor Ana Luísa Nobre Fred, Professora Associada do Instituto Superior Técnico, da Universidade de Lisboa

2014

Abstract

In the field of machine learning, the classical approach to sequence classification is based on statistical learning. This kind of problem is traditionally posed in a probabilistic framework, for which feature extraction and selection are essential to obtain the information needed to build statistical models. However, in practice, careful feature engineering and sophisticated preprocessing procedures are needed to obtain good features. Those procedures may thus become prohibitive for massive data collections. Moreover, the preprocessing is often task-specific, thus have to be redesigned and reapplied when the same data is used in a different application.

During the last decade, researchers have tried to find alternative methods that implement so-called universal classifiers, in the sense that they do not depend on prior assumptions about the unknown sequences/sources and do not require feature extraction or selection.

This thesis addresses compression-based dissimilarity measures and their use for the classification of sequences from different types of sources. We propose information theoretic measures that exploit the concept of relative entropy and a supervised classification method which use these type of measures as features in a dissimilarity space. We apply the developed methods in text classification and electrocardiographic biometrics.

Experimental results on public domain datasets show that the proposed dissimilarity measures and classification methods approximate or even outperform, in terms of accuracy, the state-of-the-art competitors in some benchmark problems.

Key-words: Machine Learning, Sequence Classification, Data Compression, Dissimilarity Space, Dissimilarity Measure, Relative Entropy, Ziv-Merhav Method, Cross-Parsing Algorithm.

Resumo

No campo da aprendizagem automática, a abordagem clássica para a classificação de sequências baseia-se em aprendizagem estatística. Estes problemas são habitualmente formulados probabilisticamente, pelo que a extração e seleção de características são essenciais para obter a informação necessária à construção de modelos estatísticos. No entanto, na prática, a obtenção de características adequadas é uma tarefa difícil habitualmente suportada em métodos de pré-processamento sofisticados. Consequentemente, estes métodos podem tornar-se proibitivos para conjuntos de dados de grande dimensão. Adicionalmente, o pré-processamento tende a ser específico de cada tarefa, pelo que precisa de ser redesenhado e reaplicado para cada aplicação diferente.

Durante a última década, tem-se tentado obter métodos alternativos que implementam métodos ditos universais, no sentido em que não dependem de hipótese prévias acerca das sequência/fontes e não requerem extracção ou selecção de características.

Esta tese estuda medidas de dissemelhança baseadas em compressão e o seu uso para classificar sequências de diferentes tipos. Propõem-se medidas de teoria da informação que exploram o conceito de entropia relativa e métodos de classificação supervisionada que usam estas medidas como características num espaço de dissemelhança. Os métodos desenvolvidos são aplicados em classificação de texto e em biometria electrocardiográfica.

Os resultados experimentais com conjuntos de dados do domínio público, para vários problemas de classificação, mostram que as medidas e os métodos propostos aproximam ou até superam, em termos de precisão, os métodos que constituem o estado da arte para alguns problemas de referência.

Palavras chave: Aprendizagem Automática, Classificação de Sequências, Compressão de Dados, Espaço de Dissemelhanças, Medidas de Dissemelhança, Entropia Relativa, Método de Ziv-Merhav, Algoritmo da Descrição Cruzada.

Acknowledgements

I would like to express my gratitude to my supervisor Prof. Mário Figueiredo, for his guidance, support and friendship during all this years of my PhD research. I am grateful for all that I learnt from him and his excellent work as a researcher that has truly inspired me. I thank Prof. Ana Fred, in collaboration with whom this work was done, for her constant stimulus and support, for the precious inputs and sound advice. Both were very important to go through this unknown path that was this research work. I would also like to thank Prof. José Bioucas Dias, for his incentive on this work and his constant availability for help.

Thanks are given to Mark Nelson and Jesper Larsson for providing source code for LZSS and to suffix trees algorithms respectively.

This work was carried out at the Instituto de Telecomunicaões (IT), Pólo de Lisboa. I thank IT for providing the office space and facilities I needed to carry out this work. I gratefully acknowledge both Instituto Politécnico de Lisboa (IPL) and Instituto Superior de Engenharia de Lisboa (ISEL) for grant SFRH/PROTEC/49615/2009, which partially supported this work from October 2009 to October 2012.

Good examples along the learning path brought me to this research work. Hence, I thank Prof. Pimenta Rodrigues, Prof. José Paraízo, Prof. José Leitão, Prof. Fernando Sousa and Prof. Arnaldo Abrantes.

For their support and encouragement along the past years, I thank my colleagues Paulo Marques, Artur Ferreira, Hugo Silva, André Lourenço, Manya Afonso, Pedro Sampaio, Maria Isabel Milho, Tiago Dias, Pedro Miguens, Nuno Leite, and my friends Vivaldo Gonçalves, João Vargues, Fernanda Delgado and Nuno Matos.

I thank the good influence over my life of my cousins Carlos Coutinho, David F. Pereira Coutinho and Arnaldo Pereira Coutinho, and my good friend António Teixeira.

And of course, I thank my parents, sister and brother for their love and encouragement.

Last but not least I owe the biggest thank to Isabel Kleinsteuber for her patience and never-ending support.

David Pereira Coutinho
Lisboa, July 2014

To my parents
José Antão and Cleópatra Pereira Coutinho

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Problem Definition

The development of global communication systems, such as the Internet and mobile networks, are making electronic information easily available for worldwide users. In the emerging context of massive collections of online information, such as e-mail messages, music files, biometric data, product reviews, and eBooks, for example, automatic data classification plays an important role, namely in the growing market of the handheld computers and smart phones applications. Let us first introduce a general notion of *classification* problems: given a set of classes, we seek to determine which class(es) a given object belongs to. This task, which may often be easily done by humans, is what we want to learn how to perform automatically with a computer. An example application, from an information retrieval context, could be automatic sentiment classification of product reviews, as positive or negative, allowing user searching for negative reviews before buying a product, to make sure it has no undesirable features or quality problems.

In machine learning, *classification* is the problem of learning how to decide to which of a set of categories (or classes) a new object belongs, given a training dataset containing objects whose category is known. The decision criterion of the classifier should be learned automatically from training data, which requires a number of good data examples (or training objects) for each class. When the learning method uses statistical information about the set of objects, the approach is called statistical learning. Individual observations are usually analyzed and a set of features extracted; given that set of features, an algorithm implements classification by mapping the input data into a category. To improve the efficiency and accuracy of the algorithms, a technique known as feature selection is commonly applied before learning and classification occur. More formally, given a feature space $\mathcal{S}$ and a set of possible categories Ω, a classifier is a mathematical function

$$\mathcal{C} : \mathcal{S} \rightarrow \Omega,$$

where each observation $x \in \mathcal{S}$ is associated with a category $\hat{w} \in \Omega$. Many different methods can be used for classification, including k-nearest neighbors (k-NN) [29], naïve Bayes [91], and support vector machines (SVM) [57] among others. Arguably the most difficult task in classification is to choose an appropriate set of features that allows machine learning algorithms to provide accurate classification. Most state-of-the-art techniques for this task involve careful feature engineering and a preprocessing stage, which are very time consuming procedures. For a comprehensive introduction on this subject see [96, 35, 9]. Despite the possible different types of individual observations, in this thesis we only address the problem of sequence classification (where each 'object' is a sequence of symbols) focusing on applications involving text and electrocardiographic data.

Text classification (or *categorization*) is the problem of assigning a text to one or more of a predefined set of classes. Examples of applications are: (i) topic classification [57], where the task is to decide which topic(s) is (are) addressed in a text; (ii) sentiment analysis (SA) [101], which is the task of automatically classifying a text, not in terms of topic, but according to the overall sentiment it expresses, e.g., determining whether a user review of some product or service is positive or negative; (iii) authorship attribution (AA) [113], where the task is to assign a text of an unknown author to one of a set of possible authors.

Classical techniques for these (and other) text classification problems are based on statistical and computational tools that require careful feature engineering and sophisticated preprocessing, which may become prohibitive due to time consumption and are specifically tailored to each application. Defining a similarity measure between texts (or, more generally, finite sequences of symbols) that allows addressing classification problems, without explicitly modeling their statistical behavior, is a fundamental problem in this context, which we address in this thesis.

The electrocardiogram (ECG) is an emerging biometric measure for which there is a strong evidence that it is sufficiently discriminative to identify individuals from a large population. In the context of this thesis, *electrocardiogram classification* is the problem of recognizing a subject from his/her ECG, in the presence of a database containing ECG data of all the system users. In 2001, Biel et al. [8] proposed a fiducial method for feature extraction, which were used for database storage and classification. *Fiducial* methods use points of interest within a single heartbeat waveform, such as local maxima or minima; that is, they use references to allow the definition of features like latency times and amplitudes [8]. On the other hand, *non-fiducial* techniques were also proposed since 2001 [17, 88], which aim at extracting discriminative information from the ECG waveform without having to extract fiducia. A global pattern from several heartbeat waveforms may be used as a feature, or wavelet or DCT coefficients are extracted and used as features (e.g., [18, 88]). We concentrate on ECG classification using a non-fiducial approach, where the first necessary step is to convert ECG samples into sequences of symbols (strings) from a 256 symbols alphabet, using 8 bit quantization. Although information is lost due the quantization process, enough discriminative information is preserved as will be shown by the experimental results. In this manner, text classification tools may be used for electrocardiogram classification.

In this thesis, we address the problem of sequence classification ignoring a priori any information about the source model, namely for text and ECG classification ignoring either the linguistic structure of the texts and the P-QRS-T complexes structure of the ECG. We aim at taking any of these sequences, apply a classification method which does not need any specific preprocessing and so handle the sequences without distinction as in universal classification [131]. Despite the fact that source model information is ignored, good classification results can be achieved, as will be shown experimentally. This 'agnostic' approach avoids preprocessing and feature selection steps, which are very time consuming and problem-specific. Because the classification system design is not tuned for a certain type of source, the tools developed with this methods will have a broader application range, as for example, an authorship attribution application developed in this fashion may be applied for text independently of its written language.

1.2 Thesis Contributions

The main contributions of this thesis are:

- an original implementation of the information theoretic dissimilarity measure proposed by Ziv-Merhav [131], which is an empirical measure of the relative entropy between individual sequences that is based on self and cross parsing algorithms. We propose an efficient cross-parsing algorithm based on the Lempel-Ziv sliding window algorithm [129] and using optimized string matching data structures (suffix trees) [65], expanding the empirical measure application domain to other type of sequences than finite-order Markovian sequences of the same size;

- a new way of using the dissimilarity measures for classification purposes. We use the dissimilarity measures as features to build a classifier in a *dissimilarity space*, where any classifier working in $\mathbb{R}^n$ can be used (i.e. k-NN or SVM). We expand the state-of-the-art, by proposing a novel supervised classification method, which uses one of the different types available of compression-based measures to make universal sequence classification.

The main attraction of these compression-based methods for classification is that they avoid the problems of explicit feature extraction and selection, thus requiring virtually no preprocessing of the input sequence. In text classification, for example, such methods do not require obtaining a representation of texts, like the bag-of-words, and the classification algorithm incorporates the quantification of textual properties.

Experiments were done on both ECG signals and text sequences from publicly available datasets. Test results further enhance the applicability of the ECG signal as a biometric trait, and confirm its biometric potential even on data acquired in unrestrained scenarios, namely when using the proposed information theoretic dissimilarity measure. Regarding text classification, results show that the Ziv-Merhav relative entropy estimation method has the potential to build accurate tools for applications like authorship attribution.

1.3 Contribution Publications

In this section, we briefly state the contributions of each of the papers included in the thesis. Paper A is the starting point for a new approach of text classification based on the information theoretic concept of relative entropy. Papers B-D concentrates on the application of the cross parsing algorithm as a similarity measure to be used in biometrics problems. Finally, papers E-G focus on a new approach for text classification using compression-based dissimilarity measures in a *dissimilarity space*.

Paper A: Coutinho, D. P. and Figueiredo, A. T. (2005). Information Theoretic Text Classification Using the Ziv-Merhav Method. In *Pattern Recognition and Image Analysis - IbPRIA 2005. Springer Berlin Heidelberg*, LNCS 3523, pages 355–362

In this contribution, we propose a new approach for text classification based on our implementation of the Ziv-Merhav method for relative entropy estimation. 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.

Paper B: Coutinho, D. P., Fred, A. L., and a.T. Figueiredo, M. (2010). One-Lead ECG-based Personal Identification Using Ziv-Merhav Cross Parsing. In *Proceedings of the 20th International Conference on Pattern Recognition - ICPR 2010*, pages 3858–3861

This paper considers the use of the Ziv-Merhav cross parsing length as a similarity measure, applied in a biometrics context. The advance of falsification technology increases security concerns and gives biometrics an important role in security solutions. The electrocardiogram (ECG) is an emerging biometric that does not need liveliness verification. There is strong evidence that ECG signals contain sufficient discriminative information to allow the identification of individuals from a large population. Most approaches rely on ECG data and the fiducia of different parts of the heartbeat waveform. However non-fiducial approaches have proved recently to be also effective, and have the advantage of not relying critically on the accurate extraction of fiducia data. In this paper, we propose a new non-fiducial ECG biometric identification method based on lossless data compression techniques, namely the Ziv-Merhav cross parsing algorithm for symbol sequences (strings). Our method relies on a string similarity measure derived from algorithmic cross complexity concept and its compression-based approximation. We present results on real data, one-lead ECG, acquired during a concentration task, from 19 healthy individuals. Our approach achieves 100% subject recognition rate despite the existence of differentiated stress states.

Paper C: Coutinho, D., Fred, A., and Figueiredo, M. (2010). Personal Identification and Authentication based on One-lead ECG using Ziv-Merhav Cross Parsing. In *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems - PRIS 2010*, pages 15–24.

In this paper, we extend our new lossless data compression based ECG biometric method for both personal identification and authentication. The ECG is an emerging biometric for which there is strong evidence that ECG signals contain sufficient discriminative information to allow the recognition of individuals from a large population. Despite most approaches relying on ECG data and the fiducia of different parts of the heartbeat waveform, we propose a non-fiducial method based on the Ziv-Merhav cross parsing algorithm for symbol sequences (strings). Our method uses a string similarity measure obtained with a lossless data compression algorithm. We present results on real data, one-lead ECG, acquired during a concentration task, from 19 healthy individuals, on which our approach achieves 100% subject identification rate and an average equal error rate of 1.1% on the authentication task.

Paper D: Coutinho, D., Silva, H., Gamboa, H., Fred, A., and Figueiredo, M. (2013). Novel fiducial and non-fiducial approaches to electrocardiogram-based biometric systems. IET Biometrics, 2(2):64–75.

This contribution to the biometric context, considers the comparison of a novel fiducial approch and our non-fiducial approach using two ECG datasets, one being a publicly available benchmark dataset. The electrocardiogram (ECG) is a non-invasive and widely used technique for cardiac electrophysiological assessment. Although the ECG has traditionally only been used for functional diagnostic and evaluation, several advances in electrophysiological sensing have made available robust signal acquisition devices, particularly suited for ambulatory conditions, widening its range of applications. In particular, recent work has shown the potential of the ECG as a biometric trait, both for human identification and authentication. This paper sets the ground for an ECG-based real-time biometric system. We describe an experimental setup and the evaluation of new fiducial and non-fiducial approaches, including data acquisition, signal processing, feature extraction and analysis, and classification methodologies, showing the applicability of the ECG as a real-time biometric. Performance evaluation was done in clinical-grade ECG recording from 51 healthy control individuals (of a publicly available benchmark dataset) as well as on data collected from 26 healthy volunteers performing computer activities without any posture or motion limitations, thus simulating a regular computer usage scenario.

Paper E: Coutinho, D. P. and Figueiredo, M. A. T. (2013). An Information Theoretic Approach to Text Sentiment Analysis. In *Proceedings of 3rd International Conference on Pattern Recognition Applications and Methods - ICPRAM 2013*, pages 577–580. SciTePress.

This paper establish a new approach for text sentiment analysis using compression-based dissimilarity measures in a *dissimilarity space*. Most approaches to text sentiment analysis rely on human generated lexicon-based feature selection methods, supervised vector-based learning methods, and other solutions that seek to capture sentiment information. Most of these methods, in order to yield acceptable accuracy, require a complex preprocessing stage and careful feature engineering. This paper introduces a coding-theoretic-based sentiment analysis method that dispenses with any text preprocessing or explicit feature engineering, but approximates state-of-the-art accuracy. By applying the *Ziv-Merhav method* to estimate the relative entropy (Kullback-Leibler divergence) and the cross parsing length from pairs of sequences of text symbols, we get information theoretic measures that make very few assumptions about the models which are assumed to have generated the sequences. Using these measures, we follow a *dissimilarity space* approach, on which we apply a standard support vector machine classifier. Experimental evaluation of the proposed approach on a text sentiment analysis problem (more specifically, movie reviews sentiment polarity classification) reveals that it approximates the previous state-of-the-art, despite being much simpler than the competing methods

Paper F: Coutinho, D. P. and Figueiredo, M. A. T. (2013). On Compression-Based Text Authorship Attribution. In *Proceedings of the 19th edition of the Portuguese Conference on Pattern Recognition - RECPAD 2013*, number 4.

In this paper we extend the use of compression-based dissimilarity measures in a *dissimilarity space* to the text authorship attribution problem. Common approaches to this problem use a bag-of-words for text data representation, which demands usually some preprocessing operations, like word stemming and stop word removal, among others, followed by a carefully designed system for feature extraction and selection. However, this actions may become very time consuming in the context of today's world where the advancement of web and social network technologies have lead to a great interest in the classification of text documents and thus computers have to handle and process massive amounts of data. In this paper, we propose an efficient method for text classification, based on information theoretic dissimilarity measures, which are used to define dissimilarity-based representations. These methods dispense with any feature design or engineering, by mapping texts into a feature space using *universal* dissimilarity measures; in this space, classical classifiers (e.g., k-nearest neighbor or support vector machines) can then be used. The experiment results using a publicly available and benchmark text corpus show that the proposed method outperforms the state-of-the-art on this authorship attribution problem, without using any preprocessing or feature engineering while defining the classifier.

Paper G: Coutinho, D. P. and Figueiredo, M. A. T. (2013). Text Classification Using Compression-based Dissimilarity Measures. *Submitted to Pattern Recognition Letters*.

This contribution compiles all the work done with our new approach to text classification using compression-based dissimilarity measures in a *dissimilarity space*. Arguably the most difficult task in text classification is to choose an appropriate set of features that allows machine learning algorithms to provide accurate classification. Most state-of-the-art techniques for this task involve careful feature engineering and a pre-processing stage, which may be to expensive in the emerging context of massive collections of electronic texts. In this paper, we propose efficient methods for text classification based on information theoretic dissimilarity measures, which are used to define dissimilarity-based representations. These methods dispense with any feature design or engineering, by mapping texts into a feature space using *universal* dissimilarity measures; in this space, classical classifiers (e.g., k-nearest neighbor or support vector machines) can then be used. The reported experimental evaluation of the proposed methods, on sentiment polarity analysis and authorship attribution problems, reveals that it approximates, sometimes even outperforms previous state-of-the-art techniques, despite being much simpler, in the sense that they do not require any text preprocessing or feature engineering.

1.4 Organization of the Thesis

Five chapters and seven appendices compose this thesis. Chapter 1 introduces the sequence classification problem and the thesis contributions. Moreover, it summarizes the contributions papers.

Chapter 2 contains a brief description about the pattern recognition system model and its building blocks, along with an overview concerning text classification and ECG Biometrics applications and previous work.

The compression-based dissimilarity measures used in the thesis are introduced and explained in Chapter 3. Details are given about the proposed implementation of the Ziv-Merhav method for relative entropy estimation.

Chapter 4 describes the proposed new method for sequence classification using compression-based dissimilarity measures, along with the experimental results obtained on text classification and ECG Biometrics applications, using publicly available datasets.

Finally, concluding remarks and some pointers for possible future work are presented in Chapter 5.

All the seven contributions publications which compose the base of this thesis are included as appendix A to G.

Chapter 2

Classification of Sequences

In this chapter, we review the problem of sequence classification using classical approaches. We start with the description of the general structure for a pattern recognition system and then we discuss the two building blocks: feature extraction and classification. We conclude this section with the description of some details about text classification and ECG biometrics, which are the focus applications in this thesis, and a brief reference to previous works is included for both applications.

2.1 Pattern Recognition System Model

The problem of automatic sequence classification arises in several application-specific contexts where the main purpose is retrieving information from sequences, such as stock market time series, DNA sequences, EEG state sequences, for which the elements order in the sequence is important. However, there are two subclasses of the sequence classification problem which are of major concern to this work: classification of texts and (sampled and) quantized analog signals, namely, the electrocardiogram. In general, sequence classification is addressed using pattern recognition (PR) systems.

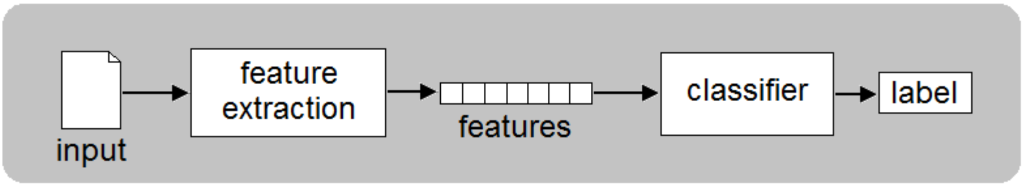


Figure 2.1: Pattern recognition system classic structure.

In machine learning, the classic structure for a *pattern recognition* system consists of two blocks, the feature extraction and the classification blocks, as depicted in Figure 2.1. Typically the first block transforms the input data into a set of features that are thought to convey the relevant information to the decision process implemented by the second block [9].

2.1.1 Feature Extraction and Selection

For most applications, the input data typically needs to be preprocessed in order to transform it into new data which will make the problem of pattern recognition easier to solve. Thus *feature extraction* takes place typically when the input data to the classifier algorithm is too large and is suspected to contain redundancy and/or irrelevancy. Then the input data will be transformed into a reduced representation: the feature vector. If the features extracted are carefully chosen then the most relevant information is expected to be extracted. This allows the classifier to perform more accurately and faster, while using this reduced representation instead of the full size input, because the analysis of a large number of variables generally requires a large amount of memory and computation power. Two popular methods for independent feature extraction are linear discriminant analysis (LDA) and principal component analysis (PCA) [120].

Besides feature extraction, another process of selecting a subset of relevant features is commonly used. This process is called *feature selection* and usually allows the replacement of a complex classifier, which uses all the extracted features, for a simpler one that uses only a subset of the initially extracted features. Feature selection techniques are a subset of the more general field of feature extraction and are often used in domains where there are many features and comparatively few data examples. Sometimes the feature vector has a large number of features (say, 10^5), and then it requires special techniques which are time efficient and computationally non-prohibitive, for example, like the recently proposed methods for feature selection on high-dimensional feature spaces by Ferreira and Figueiredo [39]. Moreover, a basic feature selection algorithm searches for the new feature subsets using an evaluation measure for scoring the different feature subsets. Some common evaluation measures are mutual information, χ^2-statistics and document frequency. A survey about feature selection methods for classification can be found in [1, 71, 33].

2.1.2 Classifier

Machine learning techniques are commonly used to build the classifier. When this approach is used, a general inductive process automatically builds a classifier by learning, given a set of previously classified objects, which are assumed to be representative of the characteristics of the classes. Once the classifier is built, it plays the most important role, as it should choose the correct class label for each input feature vector.

In basic classifiers, each input is considered individually, the classes are known a priori and only one class is assigned; other variants exist [96, 35], such as, for example, multi-label classification, where multiple classes can be assigned for each input.

For the classification process, we are given a feature vector $x \in \mathcal{S}$, where $\mathcal{S}$ is the features space, and a set of possible class labels Ω. Typically the features space $\mathcal{S}$ is a high-dimensional space and the classes $w \in \Omega$ are human defined, according the needs of an application. In order to determine a classifier, we are also given a training set $\mathcal{X}$ of labeled feature vectors (x, w), where $(x, w) \in \mathcal{S} \times \Omega$. Using a machine learning algorithm we may learn a classifier function, or rule **y**, which maps feature vectors to classes, that is

$$\hat{w} = \mathbf{y}(x)\,, \tag{2.1}$$

such that each labeled feature vector $x \in \mathcal{X}$ is assigned to a class label $\hat{w} \in \Omega$. This method of learning is called supervised learning as it is based on a training data for which the correct class labels are known. However, other methods of learning exist [9], where only a portion of the training labels are known (semi-supervised learning) or no class labels are known for any of the training examples (unsupervised learning) - clustering is the common approach to this case.

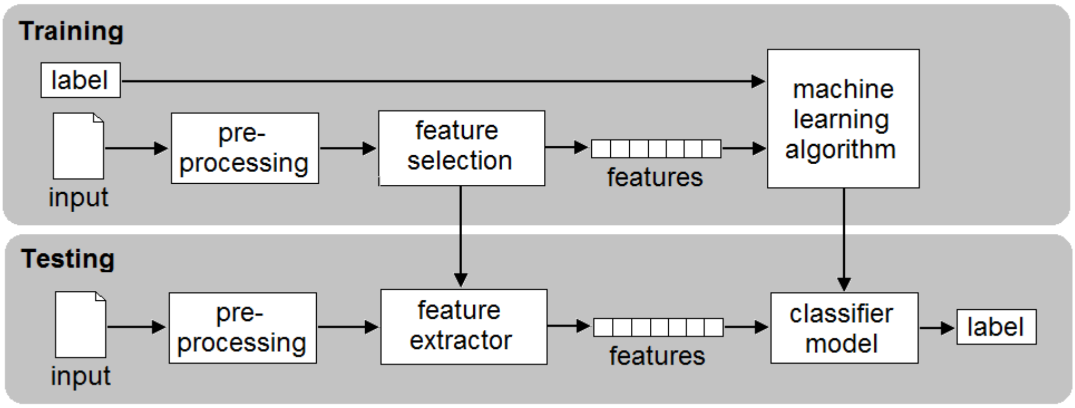


Figure 2.2: Simplified model of a supervised classifier training and testing system.

Building a supervised classifier involves two different phases, the training and testing phase, as depicted in Figure 2.2. To cope with those two phases, usually the previously classified objects are split into two sets, one for training and the other for testing. During training, a preprocessor and a feature selector (assumed here as a special type of feature extractor with the ability of selecting a subset of relevant features) are used to convert each input object to a feature vector. With these feature vectors and the respective class labels, a machine learning algorithm is fed to generate a classifier model which is translated into a classification rule. During testing, the same preprocessor and feature extractor (of the set or subset of features chosen during training) are used to convert the testing subset of input objects to feature vectors. These feature vectors are then fed into the model, which generates a predicted class label using the current learned classification rule. Given the knowledge of the correct class labels for that input object, the accuracy of the classifier can be assessed. When needed, the feature selector/extractor and the learning algorithm are modified (tuned) and the training process is repeated. Finally, when the classifier is ready to use, the very same preprocessor and feature extractor is used to convert the unseen new input objects to feature vectors. These features vectors are then fed into the classifier model, which generates a predicted class label using the learned classification rule **y** defined in (2.1).

The goal in classification is the ability to categorize correctly both testing and new objects that are different from the objects used for training, which reveals the generalization capability of the classifier. However, in practical applications, high accuracy on the training set in general does not mean that the classifier will work well on new objects, as the variability of the input objects is such that usually the training set will comprise only a very small part of the possible inputs. Thus, generalization is the central goal in pattern recognition [9].

Some typical classifiers used in pattern recognition are, for example, decision trees, neural networks, naïve Bayes (NB), k-nearest neighbor (k-NN) and support vector machines (SVM). In section 2.2.2 the NB, k-NN and SVM classifiers will be briefly described.

2.2 Text Classification

Text classification (TC), also known as text categorization, is the problem of labeling natural language texts with categories from a predefined set. Due to the availability of more powerful hardware and to application demands in the early 90s, TC become a task of major importance for the data mining, machine learning and information retrieval communities. In the present days, TC is being applied in many contexts, ranging from email spam detection, document topic classification, product review sentiment analysis, text authorship attribution, and, in general, any application requiring document handling and organization. The advantages of this approach are that it achieves an accuracy comparable to humans experts performing the same task, and makes the assignment of documents to a predefined set of categories automatically, as no human intervention is needed for the classifiers construction [99].

The most popular methods for TC are based on the machine learning paradigm, according to which a general inductive process automatically builds an automatic text classifier by learning, from a set of previously classified documents, that are assumed to be representative of the categories of interest.

There are some comprehensive reviews available in research literature related to TC [1, 99] and its applications, like authorship attribution [113] and sentiment analysis [82], where the different approaches for text representation and text classification are analyzed, while discussing the different methods used for TC in text documents at different levels (i.e. character, word, and sentence levels).

In the following sections, we will discuss briefly some detail issues concerning document representation and the most commonly used classifiers used in previous works of two specific TC applications, namely in the context of text authorship attribution and product review sentiment analysis.

2.2.1 Typical Preprocessing

Linguistic features are at the base of most text classification approaches, which require proper representation and preprocessing to obtain. The most common approach for text or document representation is called bag-of-words (BoW) and uses vectors of word frequencies, due to the fact that the simple and natural way to view a text is as a sequence of tokens grouped into sentences, each token corresponding to a word, number, or a punctuation mark. Of course, there are other alternative approaches, such as, for example, using sequences of character, or character n-grams [51] (see more alternatives in [113]). The BoW model represents each document as a bag of words or terms, where word-order is disregarded, thus loosing contextual information. Given a collection of documents D, let the vocabulary

$$V = \{t_1, t_2, ..., t_{|V|}\}$$

be the set of $|V|$ distinct words (terms) in the collection. A weight $w_{ij} > 0$ is associated with each term t_i of a document $d_j \in D$, so that

$$d_j = (w_{1j}, w_{2j}, ..., w_{|V|j}).$$

Notice that for a term t_i that does not appear in document d_j, $w_{ij} = 0$. Thus each document is represented as a vector, where each term weights is 0 or 1, for Boolean models, or is computed based on its frequency in the documents, if a generic vector space model is considered.

One of the most popular and sophisticated schemes to weigh terms is the so-called term frequency-inverse document frequency (TF-IDF), which is a numerical statistic that reflects how important a word is to a document in a collection or corpus. However, there are many other schemes to define text features [71].

One major sub-task of the text classification problem is the extraction and selection of the most appropriate features, for example, for representing the style of a text or the sentiment expressed in it. The most common features used for text representation are character and lexical features, where a text is considered as a mere sequence of characters word-tokens, respectively. Examples of character features are character types (letter, number, etc.) and character n-grams (fixed or variable length). Regarding lexical features, examples are word length, sentence length, vocabulary richness, word frequencies, etc. Notice that lexical features are more complex to obtain than character features because additional processing is needed. Other much more complex types are syntactic and semantic features exist, which require deeper linguistic analysis [113]. Feature extraction involves some text preprocessing operations, such as word-tokens extraction, erasing infrequent words, stop-word removal, stemming and computing word frequencies. This requires the availability of special tools like, for example, a tokenizer, to segment text into tokens, and a stemmer, to reduce inflected words to their base or root form.

The result of feature extraction is usually a document representation in a high-dimensional space. To improve the efficiency of the classifiers, feature selection is commonly applied in text classification.

2.2.2 Commonly Used Classifiers

In machine learning, the naïve Bayes (NB) classifier, the k-nearest neighbor (k-NN) classifier and the support vector machines (SVM) are some of the most commonly used classifiers in text classification. In order to allow the performance comparison of these classifiers, Table 2.1 shows the text classification accuracy results on Reuters-21578 corpus[1], a publicly available dataset widely used for text categorization research.

Table 2.1: Text classification accuracy results on Reuters-21578 (in percent). Results from Joachims work [57], published in 1998, while making SVM performance assessment when using bag-of-words with stemming and stop-word removal as preprocessing.

	Bayes	**k-NN**	**SVM**
earn	95.9	97.3	98.5
acq	91.5	92.0	95.2
money-fx	62.9	78.2	75.4
grain	72.5	82.2	92.4
crude	81.0	85.7	88.6
trade	50.0	77.4	76.6
interest	58.0	74.0	67.9
ship	78.7	79.2	86.0
wheat	60.6	76.6	85.2
corn	47.3	77.9	85.3
microavg.	**72.0**	**82.3**	**85.9**

Perhaps the simplest classifier that is commonly used in text classification is the probabilistic naïve Bayes (NB) classifier, which prove to be effective for many text classification problems [35]. Although in this work, we will use two other type of classifiers: k-nearest neighbor (k-NN) and support vector machines (SVM). In the experiments we will adopt k-NN classifiers due its simplicity and SVM classifiers because its efficacy, as both are some of the first choices that have been used for a long time when making text classification specific applications [1, 118, 113].

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

2.2.2.1 Naïve Bayes Classifier

The *naïve Bayes* (NB) classifier which is a simple probabilistic classifier that assumes a probability-based model for features source and applies Bayes' theorem with strong (naïve) independence assumptions, resulting in an independent feature model. The classifier combines this model with a decision rule like the *maximum a posteriori* (MAP) rule, where the most probable hypothesis is picked.

Formally, let **x** be a feature vector, from a feature space $\mathcal{S}$, with an a priori probability mass function $p(\mathbf{x})$ and a set Ω of all possible classes w. Given a new feature vector **x**, to classify it as the most probable class $\hat{w}$ from Ω, the classifier computes the maximum a posteriori probability of w, that is

$$\hat{w} = \arg\max_{w} p(w|\mathbf{x})\,. \tag{2.2}$$

Using Bayes' rule, the maximum a posteriori probability of w can be written as

$$p(w|\mathbf{x}) = \frac{p(x|w) \times p(w)}{p(\mathbf{x})} = \frac{p(\mathbf{x}, w)}{p(\mathbf{x})}\,, \tag{2.3}$$

but $p(\mathbf{x})$ plays no role in selecting $\hat{w}$, as this term is constant for all classes and thus it has no effect on the maximization of (2.2). Then the important role is played only by the joint probability $p(\mathbf{x}, w)$ which can be rewritten, using the chain rule for repeated applications of the definition of conditional probability, as follows

$$\begin{aligned} p(\mathbf{x}, w) &= p(w)\, p(x_1, ..., x_n|w) \\ &= p(w)\, p(x_1|w)\, p(x_2, ..., x_n|w, x_1) \\ &= p(w)\, p(x_1|w)\, p(x_2|w, x_1)\, p(x_3, ..., x_n|w, x_1, x_2) \\ &= p(w) \prod_{i=1}^{n} p(x_i|w, ..., x_{n-1})\,. \end{aligned} \tag{2.4}$$

Now, the naïve criteria allows to decompose the joint probability $p(\mathbf{x}, w)$ by assuming the conditional independence of features x_i given the class w, that is $p(x_i|w, ..., x_{n-1}) = p(x_i|w)$ for $\forall i$, and so

$$p(\mathbf{x}, w) = \prod_{i=1}^{n} p(x_i|w)\,. \tag{2.5}$$

Finally, given (2.3) and (2.5), replacing in (2.2) the naïve Bayes classification rule becomes

$$\hat{w} = \arg\max_{w} \prod_{i=1}^{n} p(x_i|w)\,. \tag{2.6}$$

Despite its naïve design and apparently oversimplified assumptions, NB classifiers work well in many application and its optimality has been theoretically discussed [34]. Of course, there are many other classifiers that outperform the NB classifiers in specific application such as text classification [113, 101].

2.2.2.2 k-Nearest Neighbor Classifier

The *k-nearest neighbor* (k-NN) classifier [29] is an instance-based classifier that uses similarity measures to perform the classification. The key idea is that objects which belong to the same class are likely to be similar to another (unknown) object that belongs to that same class, where the *dot product* or the *cosine metric* are some of the typical similarity measures used to determine objects similarity.

This technique relies on training data (objects) to determine the class of a new (unknown) object, therefore it does not need to build an explicit declarative model for the classes. Thus, to classify a new object we determine the k nearest neighbors in the training data and then we assign to the new object the class having the majority of representatives (votes) amongst those k neighbors, where $k \geq 1$. Notice that ties can be broken at random.

More formally, let **x** be the unknown object to be classified, let Ω be the set of all possible classes and let one of such classes be $w_i \in \Omega$. Given an object **x** and a training data set of examples $\mathcal{T}$, the first step is to find the k nearest neighbors of **x** in the training data, that is

$$KNN\left(\mathbf{x}, \mathcal{T}\right) = \left\{n_1, n_2, ..., n_k\right\} = \mathcal{N} .$$

Then, according to this technique key idea, the most probable class

$$\hat{w} = \arg\max_{w_i} \; p(w_i|\mathbf{x}) \,,$$

is the most voted class in $\mathcal{N}$, that is

$$\hat{w} = \arg\max_{w_i} \; votes(w_i, \mathcal{N}) \,.$$

Now, we only need to compute the votes for each class using

$$votes\left(w_i, \mathcal{N}\right) = \sum_{n_j \in \mathcal{N}} \delta\left(n_j, w_i\right) \,,$$

where $\delta\left(n_j, w_i\right)$ equals to 1 if n_j belongs to w_i, or 0 otherwise.

The particular case of $k = 1$ is called *nearest neighbor*, as the new object is simply assign to the class of the nearest neighbor in the training set. Despite its simplicity, the nearest neighbor ($k = 1$) classifier has an interesting property that is, in the limit $|\mathcal{T}| \rightarrow \infty$, the error rate is less than twice of the minimum error rate achievable with an optimal classifier using the true class distributions [29].

2.2.2.3 Support Vector Machines

The *support vector machines* (SVM) [21] are a statistical classification method often applied with success to text classification problems. It is a vector space based method that allows to built highly effective classifiers which can outperform the NB and k-NN classifiers [1, 118, 113, 57], particularly in situation with little training data. The basic idea of SVM classifiers is to map a sequence into a feature space and find in that space the optimal boundaries which can best separate the different classes.

Consider the example illustrated in Figure 2.3, in which we have separable training data points (examples) from two classes. There are lots of possible linear separators, although it is evident that the separator (hyperplane) which provides the best separation between the two classes, is the one where the normal distance of any of the data points (margin) from it is the largest. Therefore, that separator is the *maximum margin hyperplane* and the data points which determine the margin are called *support vectors*.

The SVM are fundamentally a two-class classifier. However, in practice, we often have to tackle problems involving more than two classes. For those cases, several methods have been proposed to reduce the single multiclass problem into multiple binary classification problems , while some other authors proposed methods that consider all classes at once. An extensive comparison of methods for multiclass SVM can be found in [52].

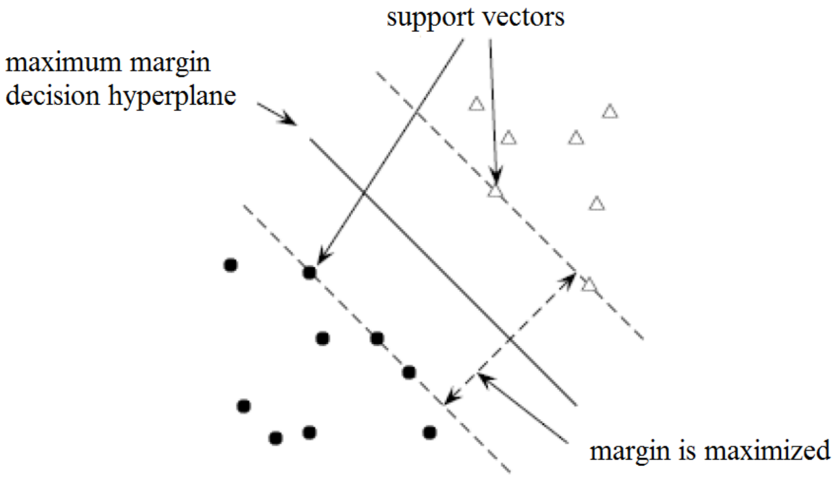


Figure 2.3: Suport vector machines find maximum margin decision hyperplane which separates the training examples of the two classes. The 5 data points closest to the hyperplane are called *support vectors* and define the margins of the classifier.

The problem of finding the maximum margin hyperplane can be set as a constrained optimization problem. Formally, given some data vectors from a feature space with dimension p and a training set $\mathcal{T}$, containing n examples $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$, the optimization of SVM (dual form) is given by

$$L(\alpha) = \arg\max_{\alpha_i} \left\{ \sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i\, \alpha_j\, y_i\, y_j\; k(\mathbf{x}_i, \mathbf{x}_j) \right\} ,$$

subject to (for all i = 1,... , n) the Lagrange multipliers be positive, $\alpha_i \geq 0$, and to the constraint

$$\sum_{i=1}^{n} \alpha_i\, y_i \;=\; 0\,,$$

where $L(\alpha) = \{\alpha_i, ..., \alpha_n\}$ is the set of Lagrange multipliers to be found and the kernel is defined by the dot product $(\cdot)$ between $\mathbf{x}_i$ and $\mathbf{x}_j$, that is $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$.

2.2.3 Authorship Attribution Previous Work

The other type of TC problem considered in this work is *authorship attribution* (AA). In the typical authorship attribution problem, a text of unknown authorship is assigned to one candidate author, given a set of candidate authors for whom text samples of undisputed authorship are available. From a machine learning standpoint, this is a multi-class case of single-label text categorization[99].

Addressing AA with statistical and computational tools has a long history, which can arguably be traced back to the seminal study on the authorship of the disputed *Federalist Papers* published by Mosteller and Wallace [76]. The Federalist Papers are a series of 85 political essays published in 1788 by Alexander Hamilton, James Madison, and John Jay, three of the so-called founding fathers of the constitution of the USA. Initially, the identity of the author of each essay was kept secret; although later the authors claimed authorship of their essays and some disputes arose. Presently, experts consider that 73 texts can be considered as having known author, while 12 are of disputed authorship. Mosteller and Wallace [76] proposed a Bayesian statistical method based on the frequencies of a small set of common words (e.g., “and”, “to”, ...), which produced good discrimination results. Since then, many other different methods have been proposed which led to great research advances in this field [61, 72, 51, 48]. But let us now focus on some particular approaches to AA, which use datasets and methods of our interest.

Benedetto et al. [6] proposed a simplified “distance” function between a pair of texts, based on the description length obtained by encoding one text using a code (a model) optimized for the other text; in practice, they propose computing this distance by concatenating the two texts and compressing the result using an off-the-shelf universal encoder, such as *gzip* or *zip*. In order to evaluate the accuracy of their method, the authors carried out experiments using a corpus of 90 texts of Italian authors (to which we will refer as the *Italian Corpus*[2], and which we will use as a benchmark dataset), reporting an accuracy of 93.3%.

[2] Available at `http://www.liberliber.it`

However, this method has some weaknesses. Namely, *gzip* is a public domain dictionary-based compressor that uses an algorithm with a sliding window of length 32 Kbytes to build the dictionary[3]; thus, if the model text is long enough, its beginning will be ignored when *gzip* is compressing its concatenation with the other text. Furthermore, if the other text is long enough, the model will disappear from the dictionary after a while. Puglisi et al. [89] studied in detail what happens when a dictionary-based compression algorithm, such as *gzip*, tries to optimize its features at the interface between two different texts.

To overcome some of the weaknesses of the approach of Benedetto et al. [6], Coutinho and Figueiredo [22] proposed a method based on an estimator of the relative entropy between pairs of sequences of symbols introduced by Ziv and Merhav [131]. An accuracy of 95.4% on the *Italian Corpus* was achieved by a NN classifier based on that relative entropy estimate as a distance measure.

An alternative distance measure based on a computable algorithmic relative complexity was proposed by Cerra and Datcu [12]. On a pre-processed version of the *Italian Corpus*, an accuracy of 97.8% was reported by those authors. This result will be used as the AA baseline result on this dataset.

Finally, very recently, Ebrahimpour et al. [38] introduced a corpus of English texts obtained from the Project Gutenberg archives[4], containing 168 short stories by seven undisputed authors from the late 19th century and early 20th century; this will be referred to as the *English Corpus*. They claimed that authors writing styles would be the key discriminant feature, since all the selected authors wrote fictional literature in English of the same genre in the same era. Two AA methods were developed, one of which uses an SVM with features based on word frequencies, involving some text preprocessing, such as for each text, stripping it of all characters except a-z and space. Experimental results revealed an accuracy of 96.4%. Moreover, the authors applied the same methods to the Federalist Papers and reported an accuracy of 97.1%.

Table 2.2 summarizes some of the important classification results with the *Italian corpus* and *English corpus*, which will be used as AA baseline results in this thesis.

Table 2.2: Reported classification accuracies (percentage of correctly classified) in the literature over some corpora of text used in this thesis.

Corpus	**Authors**	**Method**	**Accuracy** [%]
Italian	Benedetto2002 [6]	relative entropy + NN	93.3
Italian	Coutinho2005 [22]	relative entropy + NN	95.4
Italian	Cerra2009 [12]	words preproc. + relative complexity + NN	97.8
English	Ebrahimpour2013 [38]	char. and words preproc. + BoW + SVM	96.4

[3] For details see http://www.gzip.org/algorithm.txt

[4] Available at http://www.gutenberg.org

2.2.4 Sentiment Analysis Previous Work

One of the instances of the TC problem that we address in this work is *sentiment analysis* (SA). The typical sentiment analysis problem is the task of automatically classifying a text according to the overall sentiment it expresses, e.g., determining whether a user review of some product or service is positive or negative. From a machine learning point of view, this is a binary class case of single-label text categorization[99].

Starting with the seminal work of T. Joachims [57], SVM classifiers have been one of the weapons of choice when dealing with topic-based text classification. SVM classifiers typically work on vector spaces where each text is characterized by a bag-of-words (BoW). It was thus not surprising that the initial attempts at addressing text sentiment analysis (which of course is just a special type of text classification) were also based on SVM tools and BoW-type features [83]. The early work of Pang and Lee, using this type of approach, provided a strong baseline accuracy of 82.9% in a task of movie reviews sentiment polarity (binary) classification.

Since then, the movie review dataset (also known as the sentiment polarity dataset) used in [83, 81] has become a benchmark for many sentiment classification studies. We now recall some of the best result to date on this dataset.

Whitelaw and collaborators [123], reported an accuracy of 90.2%. Their method is based on so-called *appraisal groups*, which are defined as coherent groups of words around adjectives that together express a particular opinion, such as "very funny" or "not terribly surprising". After building an apraisal lexicon (manually verified) it uses a combination of different types of appraisal group features and BoW features for training an SVM classifier.

The state-of-the-art accuracy was established by Matsumoto and collaborators [73]. They proposed a method where information about word order and syntactic relations between words in a sentence is used for training a classifier. Thus using the extracted word sub-sequences and dependency sub-trees as features for SVM training, they attained an accuracy of 93.7%.

More recently Yessenalina and colleagues [128] proposed a supervised multi-level structured model based on SVM, which learns to jointly predict the document label and the labels of a sentence subset that best explain the document sentiment. The authors treated the sentence-level labels as hidden variables so the proposed model does not required sentence-level annotation for training, avoiding this way the lower-level labellings cost. They formulate the training objective to directly optimize the document-level accuracy. This multi-level structured model achieved 93.22% document-level sentiment classification accuracy on the movie review dataset.

These and other results with corresponding references are summarized in Table 2.3. Further examples can be found in [101, 118], but all usually involving complex preprocessing stages and careful feature engineering.

Table 2.3: Baseline and best classification accuracies (percentage of correctly classified) reported in the literature over the same collection of movie reviews.

Method	**Accuracy** [%]
Pang2002 [83]	82.9
Pang2004 [81]	87.2
Whitelaw2005 [123]	90.2
Matsumoto2005 [73]	93.7
Kennedy2006 [58]	86,2
Yessenalina2010 [128]	93.2
Maas2011 [70]	88,9
Duric2011 [36]	87,5

2.3 ECG Biometrics

Biometrics deals with recognition of individuals based on their physiological or behavioral characteristics [55] and plays an important role namely in security systems. Typically there are two different recognition procedures: identification and authentication. From a machine learning perspective, the identification procedure is a multi-class case of single-label classification, where the problem is to assign an identity from a set of known subjects when we are given an unknown biometric sample. On the other hand the authentication procedure is a binary class case of single-label classification, where the problem is to confirm if the subject is who claims to be, given an unknown biometric sample and a claimed identity.

Traditional methods of biometric recognition, such as those using fingerprints or iris, provide accurate recognition but lack robustness against falsification. The electrocardiogram (ECG) is an emerging biometric tool exploiting a physiological feature that exists in all humans; there is a strong evidence that the ECG is sufficiently discriminative to identify individuals in a large population [8]. The ECG has intrinsic liveliness verification, and beyond personal identification and authentication it allows detection of different stress or emotional states [75].

In the context of this work, ECG data classification is the problem of subject recognition given ECG samples from the unknown subject, in the presence of ECG samples from all known subjects previously stored in a database.

A typical ECG signal of a normal heartbeat can be divided into 3 parts, as depicted in Figure 2.4: the P wave (or P complex), which indicates the start and end of the atrial depolarization of the heart; the QRS complex, which corresponds to the ventricular depolarization; and, finally, the T wave (or T complex), which indicates the ventricular repolarization. It is known that the shape of these complexes differs from person to person, a fact which has stimulated the use of the ECG as a biometric [8].

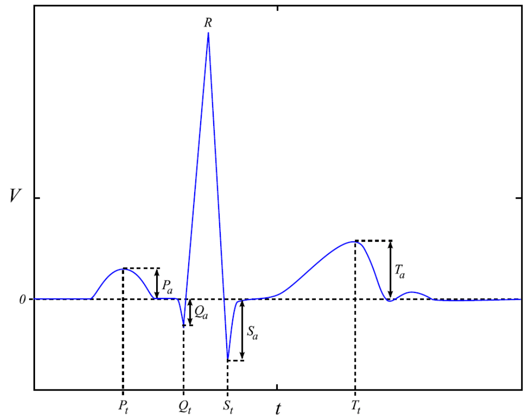


Figure 2.4: Example of four latency times (features) measured from the P, QRS and T complexes of an ECG heartbeat for feature extraction based on fiducia.

2.3.1 Fiducial and Nonfiducial Approaches

In a broad sense, one can say there are two different approaches in the literature concerning feature extraction from the ECG: fiducial [8, 104, 53], and non-fiducial [16]. *Fiducial* methods use points of interest within a single heartbeat waveform, such as local maxima or minima (e.g. P,Q,R,S,T points); these points are used as reference to allow the definition of several time and amplitude features (see Figure 2.4). *Non-fiducial* techniques extract discriminative information from the ECG waveform without localizing fiducial points. In this case, a global pattern from several heartbeat waveforms may be used as a feature. Some methods which combine these two approaches are called partially fiducial [121] (e.g., they use only the *R* peak as a reference for segmentation of the heartbeat waveforms).

2.3.2 Previous Work

Biel et al. [8] pioneered the use of the ECG as a biometric for personal identification. They initially used a 12-lead ECG, which requires meticulous and unpractical placement of the electrodes on each person, but ended up concluding that one lead was enough. Using a proprietary equipment from SIEMENS, 30 fiducial features were extracted; a feature selection algorithm allowed concluding that the best results were obtained with 10 of these features, based on *principal component analysis* (PCA) of each class. The purpose was to identify 20 subjects at rest, a task on which they achieved 100% accuracy.

Using a predetermined group of 20 subjects, selected from the MIT-BIH ECG database[5], Shen et al. performed experiments targeting ECG-based identity verification [104]. Through a template matching technique, a 95% accuracy was obtained, whereas using a *neural network* classifier lead to 80% accuracy; a method combining both techniques achieved 100%

[5] Available at http://ecg.mit.edu

identity verification accuracy.

Table 2.4: Previous work accuracy (**Accur.** stands for % of correctly classified) where the values shown are the reported results for person identification, on databases of different sizes with different number of subjects (**Subjs.**)

Ref.	**Feature**	**Method**	**Subjs.**	**Accur.** [%]
Biel2001 [8]	Fiducial	PCA	20	100
Shen2002 [104]	Fiducial	Template matching + Neural Net.	20	100
Israel2005 [53]	Fiducial	LDA	29	100
Wang2008 [121]	Non-fiducial	AC/DCT+k-NN	13	97.8
Chiu2008 [18]	Non-fiducial	Wavelet Distance	35	100
Chan2008 [16]	Non-fiducial	(fingers) Wavelet Distance	50	89

Another important step was accomplished by Israel et al. in [53]. Experiments were performed on data collected from 29 subjects while performing a set of 7 activities. The recordings were performed on the chest and neck, and 12 temporal features extracted from the signal were used. Using standard linear discriminant analysis (LDA), individual waveforms are classified and mapped to the identity of the subject by a majority voting scheme. The authors report an accuracy of 100% in subject identification.

A new method for feature extraction from the one-lead ECG signal, based on a combination of *autocorrelation analysis* (AC) with the *discrete cosine transform* (DCT), was introduced by Wang et al. in [121]. This method does not require segmentation of the ECG signal into heartbeats, with only the *R* peak detection being needed for the *QRS* window identification. In a subject identification task (on a subset of 13 subjects from the MIT-BIH dataset), the authors used a NN classifier based based on the normalized Euclidean distance between feature vectors, and reported a recognition rate of 97.8%.

A system based on a 3 step feature extraction method was introduced by Chiu et al. in [18]. The method uses the *QRS-complex* detection algorithm proposed in [110], and the *discrete wavelet transform* to extract signal features. A k-NN classifier based on the Euclidean distance between feature vectors is used. On 35 subjects from the QT database [6], the authors report an accuracy of 100% for the identification task and verification rates of 0.83% of false acceptance rate (FAR) and 0.86% of false rejection rate (FRR).

Human identification based on an ECG acquired from the fingers is possible, as shown by Chan and Hamdy in [16]. The authors introduced a simple non-clinical data acquisition setup based on 2 button electrodes, which the subjects hold between the pads of their thumb and index fingers. The *P-QRS-T complexes* are detected and temporally aligned, in order to

[6] Available at http://www.physionet.org/physiobank/database/qtdb

compute an average ECG; the proposed classifier uses a distance measure based on wavelet coefficients. A 89% identification accuracy was achieved on a dataset of 50 individuals.

Table 2.4 summarizes the main characteristics and results of the several approaches reviewed in the previous paragraphs; more details on each method may of course be found on the corresponding publications.

Chapter 3

Compression-Based Dissimilarity Measures

Much work has been done in recent years concerning the design and development of information theoretic dissimilarity measures [131, 6, 116, 67, 22, 12, 119, 13, 50, 23, 24, 14]. Most of the proposed approaches were developed in a Kolmogorov complexity context, but an alternative approach can be based on Shannon's relative entropy, as we will show. We use the following notation: let Σ be a finite alphabet; let $\mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{y} = (y_1, y_2, ..., y_m)$ be two sequences of symbols (here also termed *strings*) from Σ, respectively with length of n and m symbols; assume that $()$ is the empty sequence and that $|\mathbf{x}|$ denotes the length of $\mathbf{x}$; the subsequence of $\mathbf{x}$ between positions i and j is denoted as $x[i, j] = (x_i, ..., x_j)$; finally, $\mathbf{x} \circ \mathbf{y} = (x_1, x_2, ..., x_n, y_1, y_2, ..., y_m)$ denotes the concatenation of $\mathbf{x}$ and $\mathbf{y}$.

3.1 Contributions

In this thesis, we present an original implementation of the information theoretic dissimilarity measure proposed by Ziv-Merhav [131], which is an empirical measure of the relative entropy between individual sequences that is based on self and cross parsing algorithms. We propose an efficient cross-parsing algorithm based on the Lempel-Ziv sliding window algorithm [129] and using optimized string matching data structures (suffix trees) [65], expanding the empirical measure application domain to other type of sequences than finite-order Markovian sequences of the same size.

3.2 Introduction

Classical information theory has its origin in the work of C.E. Shannon, who published in 1948 the seminal paper [100]. He introduced the concept of *entropy*, which is a measure of the uncertainty about the outcomes of a random variable. Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = P(X = x)$, $x \in \mathcal{X}$. The *entropy* $H(X)$ of a discrete random variable is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \, \log_2 p(x) \, . \tag{3.1}$$

As is famously known, this quantity can be interpreted as a measure of the average amount of information (expressed in *bits*), needed to describe the random variable X. Thus, one *bit* is the amount of information needed to describe, on average, the outcome of a uniformly distributed binary random variable (e.g., a fair coin). Also, $H(X)$ can be seen as the shortest expected length with which it is possible to encode losslessly the outcomes of the random variable X.

So, as a first approach we can describe the information source behavior with the random variable X. This is the standard model for a memoryless source while producing independent symbols. A more realistic source model takes into account the dependency of each symbol on the k previous ones (memory). Using a Markov chain as proposed in Shannon seminal paper[100], that is,

$$p(x_n | x_{n-1}, x_{n-2}, ..., x_{n-k}), \tag{3.2}$$

denotes the transition probabilities of the Markov chain which reflects the memory effect of the source. In this case, rather than the entropy as given in (3.1), the average uncertainty is measured by the so-called *entropy rate*:

$$H'(X) = \lim_{n \to \infty} H(X_n | X_{n-1}, X_{n-2}, ..., X_{n-k}) \, . \tag{3.3}$$

However, this entropy rate definition relies on knowledge of the transition probabilities which are statistically hard to estimate for Markov chains of large order (large k), and the problem of how to choose the order k also arises. Moreover, entropy is an ensemble/average quantity based on probabilistic assumptions; consequently, it does not provide the informational content of individual objects.

In contrast, given an instance x (typically a sequence of symbols or *string* **x**) its so-called *Kolmogorov complexity* $K(x)$, one of the central concepts of algorithmic information theory, is a measure of its intrinsic complexity [68]. Algorithmic information theory has its roots in the seminal work of Solomonoff, Kolmogorov, and Chaitin in the 1960s [111, 112, 60, 15], and includes several quantities defined on sequences of symbols, including complexity, randomness, and information.

The *Kolmogorov complexity* $K(x)$, or algorithmic complexity, of an instance x, is a measure of the computational resources needed to describe that instance; more precisely, it is the length (usually in bits) of the shortest possible program used as input by a universal Turing machine to produce the instance x and halt [68]. It can be shown that the Kolmogorov complexity of any sequence cannot be more than a few bytes larger than the length of the sequence itself, while low complexity sequences may have considerably shorter program descriptions. One interpretation of $K(x)$ is the quantity of information needed to recover x from scratch. However, it is known that $K(x)$ is non-computable [68], thus approximations must be used, such as the length $C(x)$ of a compressed version of x using some off-the-shelf lossless compressor.

There is a formal link between Shannon entropy and algorithmic complexity [68], as stated by the following theorem: the sum of the expected Kolmogorov complexities of all the instances x which are output of a random source X, weighted by their probabilities $p(x)$, equals the statistical Shannon entropy of X, up to an additive constant, that is

$$H(X) \leq \sum_{x} p(x)\, K(x) \ \leq H(X) + K(p) + O(1)\,, \tag{3.4}$$

where $K(p)$ denotes the so-called probability function complexity. For low complexity distributions, the impact of $K(p)$ is lower and the expected Kolmogorov complexity is close to the Shannon entropy.

Although we have seen that the memoryless model is too simplistic, to explain more simply the idea of how to measure the similarity between two sequences, we will use sequences from memoryless binary sources. So, let us now consider two memoryless sources $\mathcal{A}$ and $\mathcal{B}$ producing sequences of binary symbols; source $\mathcal{A}$ emits 0 with probability p (thus 1 with probability $1-p$), while $\mathcal{B}$ emits 0 with probability q. According to Shannon's information theory [100, 30], there are lossless compression algorithms that, applied to sequences emitted by $\mathcal{A}$, are able to encode them with an average number of bits per symbol asymptotically equal to the source entropy $H(\mathcal{A})$,

$$H(\mathcal{A}) = -p\,\log_2 p - (1-p)\,\log_2(1-p) \quad \textit{bits}/\text{symbol}. \tag{3.5}$$

An optimal code for $\mathcal{B}$ will not be optimal for $\mathcal{A}$ (unless, of course, $p=q$). The average number of excess bits per symbol that are wasted when we encode sequences emitted by $\mathcal{A}$ using an optimal code for $\mathcal{B}$ is given by the *relative entropy*, or *Kullback-Leibler* (KL) *divergence*, between the corresponding distributions [30], that is

$$D(\mathcal{A}||\mathcal{B}) = p \log_2 \frac{p}{q} + (1-p)\log_2 \frac{1-p}{1-q} \geq 0\,, \tag{3.6}$$

where the non-negativity of $D(\mathcal{A}||\mathcal{B})$ is the so-called *fundamental inequality of information*.

This fact suggests the following strategy 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 $D(\mathcal{A}||\mathcal{B})$. More precisely, let $C_{COMP}(\mathbf{x})$ be the length of the compressed sequence $\mathbf{x}$ given some off-the-shelf lossless compressor *COMP*. Then, an approximation to the relative entropy of sequence $\mathbf{y}$ with respect to sequence $\mathbf{x}$, here denoted by $H_{COMP}(\mathbf{y}||\,\mathbf{x})$, may be defined as the properly normalized difference $(C_{COMP}(\mathbf{x}\circ\mathbf{y}) - C_{COMP}(\mathbf{x}))/|\mathbf{y}|$, where $\mathbf{x}\circ\mathbf{y}$ is

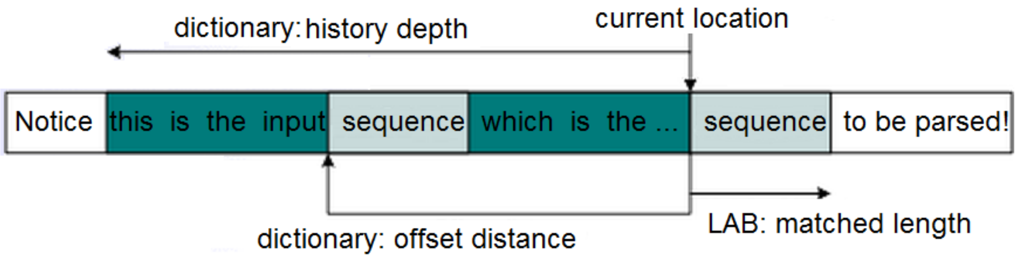


Figure 3.1: *Sliding window buffer* with the *dictionary* and *look ahead buffer* (LAB) of the lossless data compression algorithm LZ77.

the concatenation of sequences **x** and **y**. This is the rationale underlying the relative entropy estimation methods proposed by Ziv and Merhav [131], Benedetto et al. [6] and Khmelev and Teahan [59]. Notice that, the information theoretic definition of entropy by Kolmogorov motivates the use of lossless compression algorithms for estimating entropy. According to (3.4), the entropy of $\mathcal{A}$ itself can be estimated by measuring the average code length of an optimal code. This naturally leads to the idea of applying lossless compression algorithms to estimate the relative entropy.

Furthermore, since the original definition of entropy is given in terms of Markov chains of large order (by (3.3) assuming that the source has memory), a direct estimate of relative entropy can be obtained using a Markov chain of some order k on the symbols of the sequence **x** as a model for the source. The transition probabilities for Markov Chains in 3.2 are estimated from **x**, and $H_{MC}(\mathbf{y}||\,\mathbf{x})$ is defined to be the properly normalized logarithm of the probability of the sequence **y** with respect to estimated probabilities from **x**. For sequences of large size the estimate works well, but for shorter sequences it is useful to combine the probabilities of symbols using Markov chains of several orders, which are statistically hard to estimate.

Thus, to use this idea for general sources (memoryless and with memory), 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) [30].

The well-known LZ77 and LZ78 are two seminal *universal lossless compression algorithms*, introduced by Ziv and Lempel in 1977 and 1978, respectively [129, 130]. We now briefly describe LZ77, which is particularly simple and has become popular as one of the standard algorithm for lossless compression of computer files, due to its speed and efficiency. The *LZ77 algorithm* observes the input sequence through a sliding window buffer as shown in Figure 3.1. The *sliding window buffer* consist of a *dictionary* and a *look ahead buffer* (LAB). The dictionary holds the symbols already analyzed and the LAB the next symbols to be analyzed. At each step, the algorithm tries to express the sequence in the LAB as a sub-sequence 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.

3.3 The Normalized Compression Distance

One of the best known compression-based dissimilarity measure for text is the Normalized Compression Distance (NCD), proposed by Li et al. in 2004 [67]. NCD approximates the non-computable *Kolmogorov complexity* of a string **x** by the length of a compressed version of **x**, using off-the-shelf compressors, such as *gzip* or *bzip2*, and it is defined for any pair of strings **x** and **y** as

$$NCD(\mathbf{x}, \mathbf{y}) = \frac{C(\mathbf{x} \circ \mathbf{y}) - \min\{C(\mathbf{x}), C(\mathbf{y})\}}{\max\{C(\mathbf{x}), C(\mathbf{y})\}}, \tag{3.7}$$

where $C(\mathbf{x})$ is the length of string **x** after being compressed by a lossless compression algorithm, and $C(\mathbf{x} \circ \mathbf{y})$ stands for the length after compression of the concatenation of strings **x** and **y**. The authors demonstrate that it is a metric and claim that it minorizes every computable distance in the class. NCD ranges from 0 to 1 + ϵ, where 0 corresponds to **x** and **y** being identical, and 1 means maximum dissimilarity; the constant ϵ is an upper bound due to imperfections in the compression algorithms, but is unlikely to be above 0.1 for most standard compressors [67]. NCD was successfully used in clustering applications [20] despite the fact that some standard lossless compression algorithms, such as LZ77, LZ78, and even PPM (*prediction by partial matching*), are not guaranteed to satisfy these bounds.

3.4 The Ziv-Merhav Relative Entropy Estimate

In 1993 Ziv and Merhav introduced a method for measuring *relative entropy* between pairs of finite-order Markovian sequences of symbols, which can be used as a dissimilarity measure for *universal* classification [131]. The method is based on the incremental Lempel-Ziv (LZ) parsing algorithm [130] and on a variation thereof, known as *cross-parsing*.

The *incremental LZ parsing* algorithm is a self-parsing procedure of a (length n) sequence **z** into $c(\mathbf{z})$ *distinct phrases*, such that each phrase is the shortest sequence that is not a previously parsed phrase. For example, with $n = 11$ and $\mathbf{z} = (01111000110)$, the self incremental parsing yields $\{0, 1, 11, 10, 00, 110\}$, thus $c(\mathbf{z}) = 6$. Ziv and Merhav also defined in [131] a *cross-parsing* algorithm that is the sequential parsing of a sequence **z** with respect to another sequence **x**. In this case, $c(\mathbf{z}|\mathbf{x})$ denotes the number of phrases in **z** with respect to **x**. For example, with **z** as above and $\mathbf{x} = (10010100110)$, parsing **z** with respect to **x** yields the set of phrases $\{011, 110, 00110\}$, thus $c(\mathbf{z}|\mathbf{x}) = 3$. Combining these two algorithms, Ziv and Merhav proposed an estimate of the relative entropy between two ergodic sources producing the sequences **z** and **x**, which can be used as a dissimilarity measure between those sequences [131]. Specifically, they proved that for two finite order (of any order) Markovian sequences of length n the quantity

$$\Delta(\mathbf{z}||\mathbf{x}) = \frac{1}{n}\left[\, c(\mathbf{z}|\mathbf{x})\ \log_2 n - c(\mathbf{z})\ \log_2 c(\mathbf{z})\,\right] \tag{3.8}$$

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 that $(1/n)\, c(\mathbf{z})\ \log_2 c(\mathbf{z})$ is the measure of the complexity of the sequence **z** obtained by self-parsing, thus providing an estimate of its entropy according to (3.4), 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**. The difference between the two quantities does provide a measure of how different the distributions that produced the two sequences are.

3.5 The Cross-Parsing Distance

The use of the Ziv-Merhav relative entropy estimate, defined by (3.8), is not directly applicable in some scenarios, namely because it is defined for sequences of the same length n. When generalizing this definition to sequences of different lengths, several problems arise, with the size of the "model" sequence **x** having a significant impact [50]. To overcome this difficulty, Helmer et al. [50] recently introduced the *cross-parsing distance* (CPD), which is a semi-metric (i.e., of all the conditions that have to be satisfied by a metric, it only does not satisfy the *triangle inequality*) defined for any pair of sequences of symbols (strings) **x** and **y**, of length respectively $|\mathbf{x}|$ and $|\mathbf{y}|$, as

$$\text{dist}_{CPD}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\left(\frac{|s(\mathbf{x}|\mathbf{y}) \setminus \{\,\mathbf{y}\}\,|}{|\mathbf{x}|} + \frac{|s(\mathbf{y}|\mathbf{x}) \setminus \{\,\mathbf{x}\}\,|}{|\mathbf{y}|}\right), \tag{3.9}$$

where $s(\mathbf{x}|\mathbf{y})$ denotes the multiset of all phrases resulting from the cross-parsing of **x** with respect to **y** and $s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\}$ denotes the removal of a single instance of **y** from the multiset $s(\mathbf{x}|\mathbf{y})$ (if one exists, even if multiple copies exist). If the first not yet parsed symbol in **x** is not found in **y**, then the parsing is simply the symbol itself. For example, if $\mathbf{x} = (ababacbaba)$ and $\mathbf{y} = (aba)$, then $s(\mathbf{x}|\mathbf{y}) = \{aba, ba, c, ba, ba\}$, $s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\} = \{ba, c, ba, ba\}$, and $|s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\}| = 4$. Notice that if $\mathbf{x} = \mathbf{y}$, then $s(\mathbf{x}|\mathbf{y}) = \{\mathbf{y}\}$ and $s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\} = \emptyset$, thus $|s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\}| = 0$; of course, the reciprocal is true, thus $\mathbf{x} = \mathbf{y}$ implies that $\text{dist}_{CPD}(\mathbf{x}, \mathbf{y}) = 0$. In contrast, if no symbol in **x** can be found in **y**, then $s(\mathbf{x}|\mathbf{y}) = \{x_1, x_2, ..., x_{|\mathbf{x}|}\}$ (where x_i denotes the i-th symbol of string **x**), thus $s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\} = s(\mathbf{x}|\mathbf{y})$ and $|s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\}| = |\mathbf{x}|$; consequently, in this case, $\text{dist}_{CPD}(\mathbf{x}, \mathbf{y}) = 1$.

However, we had previously proposed and used successfully the cross parsing length as a dissimilarity measure in an ECG-based biometric recognition problem[26]; it was a normalized cross-parsing function defined as $\frac{c(\mathbf{x}|\mathbf{y})}{|\mathbf{x}|}$, where $c(\mathbf{x}|\mathbf{y})$ denote the number of phrases resulting from the cross-parsing of **x** with respect to **y** (as included in 3.8). Thus, now we propose to use a variant of CPD, a modified version of the dist_{CPD} defined by (3.9), which we call ***CPdist*** and define as

$$\text{CP}_{dist}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\left(\frac{c(\mathbf{x}|\mathbf{y}) - 1_{\mathbf{x}=\mathbf{y}}}{|\mathbf{x}|} + \frac{c(\mathbf{y}|\mathbf{x}) - 1_{\mathbf{y}=\mathbf{x}}}{|\mathbf{y}|}\right), \tag{3.10}$$

where $c(\mathbf{x}|\mathbf{y})$ is the number of phrases resulting from the cross-parsing and 1_A is the indicator function of the proposition A.

3.6 Proposed Ziv-Merhav Method Implementations

At the core of the Ziv and Merhav method [131] for measuring relative entropy between pairs of finite-order Markovian sequences proposed by is the cross-parsing algorithm. In 2005 we published the work [22] where we introduce the first known implementation of the cross-parsing algorithm based on the LZ77 algorithm (here termed CP77). Figure 3.2 depicts the original LZ77 sliding window and our modified LZ77 sliding window implementation for cross-parsing, where the dictionary is *static* and only the look ahead buffer (LAB) slides over the input sequence.

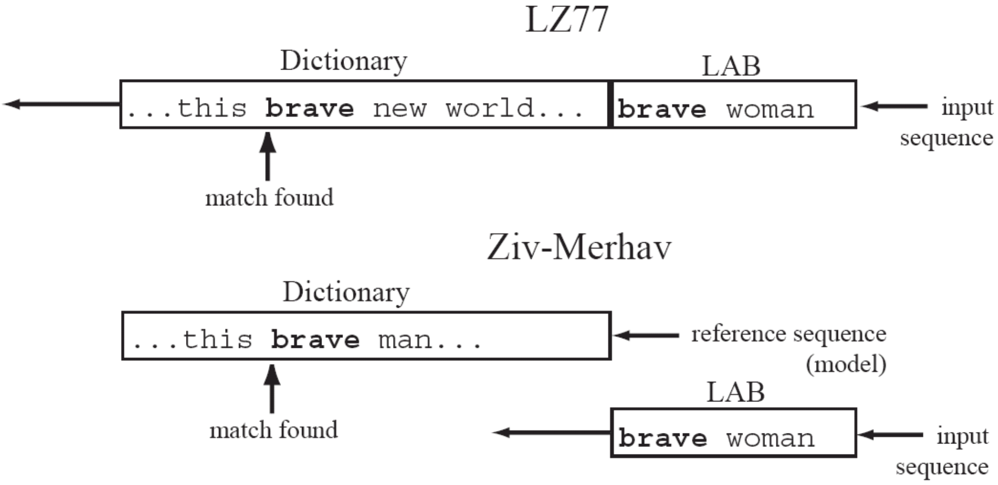


Figure 3.2: The original LZ77 sliding window and our modified implementation for cross-parsing, where the *dictionary* is static and only the *look ahead buffer* (LAB) slides over the input sequence.

Let us recall the LZ77 algorithm *self-parsing procedure* of the sequence **x** with length n: initialize a dictionary to the alphabet Σ; assume to have encoded $x[1,i]$; let **s** be the longest prefix of $x[i+1,n]$ found in the LAB that has an occurrence in the dictionary starting at some offset $j \leq i$, with length $|\mathbf{s}|$, and let $x[i+|\mathbf{s}|+1] = a$ be the innovation symbol; then, append to the encoding the dictionary reference, that is the triplet $\langle j, |\mathbf{s}|, a\rangle$, and repeat the process starting at $x[i+|\mathbf{s}|+2]$. Optimal encoding of the triplet $\langle j, |\mathbf{s}|, a\rangle$ requires a number of binary digits which is roughly $(log_2(i) + log_2(n) + log_2(|\Sigma|))$, and that may result in compression at the end. Furthermore, this procedure implementation takes $O(n)$ steps.

In 1982 Storer and Szymanski [114] proposed an efficient implementation of the LZ77 lossless algorithm, named LZSS. It is a lossless data compression algorithm which is a modified version of LZ77, with the following main differences: (i) a one-bit flag is used to indicate whether the next encoded prefix is a literal (symbol) or a dictionary reference; (ii) if the length of the encoded prefix is less than a "break even" threshold, its symbols are encoded as literals. So, basically, triplets $\langle j, |\mathbf{s}|, a\rangle$ are replaced by pairs $\langle j, |\mathbf{s}|\rangle$, and innovation symbols $\langle a\rangle$ are encoded as literals [78]. Notice that according to Ziv and Merhav [131], counting the number of pairs and literals, which is the number of distinct parsing phrases $c(\mathbf{x})$, is enough for relative entropy estimation and so the encoding process will be avoided in this method implementation.

An important issue about these data compression algorithms implementation is the way prefix search is made. Efficient prefix search is obtained by using a *suffix tree* [122], which is a data structure that stores all the different suffixes of a string in a way that allows for (fast) substring search in linear time (allows checking if **s** is a substring in $O(|\mathbf{s}|)$ time). Moreover, building a suffix tree for a given string in linear time (i.e., $O(|\mathbf{x}|)$) is also possible using the algorithm proposed by Ukkonen in [117]. This is an online algorithm where the suffix tree is constructed on the fly while parsing the string. When moving from symbol x_i to x_{i+1} in a string **x** during parsing, all the suffixes for the string from x_1 to x_i already stored in the tree are extended by x_{i+1}.

In this thesis we introduce two different implementations for the cross-parsing algorithm: the Static Cross-Parsing Algorithm (CP77) and the Incremental Cross-Parsing Algorithm (CP77inc). Both are inspired on the LZSS implementation of Mark Nelson in [78] and use the suffix-tree-based sliding window code provided by Larsson in [65], with a 2 Mbyte sliding window (dictionary) and a 256 byte look ahead buffer. The Static Cross-Parsing Algorithm uses a static dictionary built from the model sequence at the beginning of the algorithm before the parsing of the 'unknown' sequence starts (see more details in [22]). On the other hand the Incremental Cross-Parsing Algorithm uses an adaptive dictionary that is incrementally updated based on parsing of the model sequence and as the parsing of the 'unknown' sequence goes on.

Algorithm 1 illustrates how the proposed *incremental version of* CP77, here termed CP77inc, uses a suffix tree with an LZSS-based algorithm to incrementally cross-parse strings in linear time. Notice that the difference between CP77 and CP77inc is that the latter uses an *adaptive dictionary* instead of a static one. Thus, CP77 is simpler as it only needs to copy the model sequence to the dictionary at the beginning of the algorithm (the dictionary is updated only once); in contrast, CP77inc updates it every time a literal or a prefix is found.

The cross-parsing of string **z** with respect to the string **x** involves some details, which we

Algorithm 1 CP77inc: Incremental Cross-Parsing Procedure

```
Input: $z$: $1 \times n$ vector, containing the unknown sequence with $n$ symbols.
       $x$: $1 \times m$ vector, containing the model sequence with $m$ symbols.
       WINDOWSIZE: an integer constant, setting the sliding window size.
       LOOKAHEADSIZE: an integer constant, setting the size for the look ahead buffer (lab).
Output: $c_{zx}$: an integer, denoting the number of phrases in $z$ with respect to $x$.

initialize suffix tree based sliding window with WINDOWSIZE ;
$lab_z \leftarrow z[1, \text{LOOKAHEADSIZE}]$, $lab_x \leftarrow x[1, \text{LOOKAHEADSIZE}]$ ;
$i \leftarrow 1$ , $j \leftarrow 1$ , $c_{zx} \leftarrow 0$ ;
$Dx \leftarrow ()$ ;                                              { // empty dictionary }
while $i \leq |z|$ do
    $c_{zx} \leftarrow c_{zx} + 1$;
    { // cross-parsing of $z$ given $x$ }
    find prefix with largest $len$ so that $lab_z[1, len]$ can be found in $Dx$ ;
    if match not found then
        $len \leftarrow 1$ ;                                      { // literal }
    end if
    $lab_z \leftarrow \text{updateLAB}(z, i, len)$;
    $i \leftarrow i + len$ ;
    { // dictionary update using self parsing over $x$ }
    if $j \leq |x|$ then
        find prefix with largest $len$ so that $lab_x[1, len]$ can be found in $Dx$ ;
        if match not found then
            $len \leftarrow 1$ ;
        end if
        $Dx \leftarrow Dx \circ x[j, j + len - 1]$ ;             { // append match to dictionary }
        add to suffix tree the prefix found $x[j, j + len - 1]$ ;
        $lab_x \leftarrow \text{updateLAB}(x, j, len)$ ;
        $j \leftarrow j + len$ ;
    end if
end while
return $c_{zx}$ ;
```

now briefly describe; it uses one sliding window to hold both the dictionary D_x and the look ahead buffer lab_x of the model string **x**. In addition, it uses another sliding window (smaller) to hold the look ahead buffer lab_z for the unknown string **z**. Notice that the dictionary D_x is empty at the beginning. Then, a two-step loop is repeated until the end of **z** is reached: the cross-parsing of **z** given **x**; and the self parsing of **x** including dictionary update as long as **x** lasts. This makes sequences of different lengths allowed, by stopping the dictionary update whenever the end of **x** is reached and keep using it as a “static” dictionary. Every time the loop is executed a counter c_{zx} is incremented.

Finally, the estimating method of relative entropy via definition (3.8) will be called as ***ZMM*** when based on CP77, while the method that uses the proposed algorithm CP77inc will be called ***ZMMinc***.

Chapter 4

Classification Using Dissimilarity Measures

In this chapter we will discuss the goal, the advantages and drawbacks of a novel approach for classification using dissimilarity measures. We explain how is applied compression-based dissimilarity measures in some specific classification problems, such as text authorship attribution and sentiment analysis, as well as in the apparently unrelated problem of ECG biometrics.

4.1 Contributions

Along with the proposal of new information theoretic dissimilarity measures, we also introduce in this thesis a new way of using the dissimilarity measures for classification purposes. We use the dissimilarity measures as features to build a classifier in a *dissimilarity space*, where any classifier working in $\mathbb{R}^n$ can be used (i.e. k-NN or SVM). We expand the state-of-the-art, by proposing a novel supervised classification method, which use one of the different types available of compression-based measures to make universal sequence classification.

4.2 Introduction

The problems of text classification and ECG recognition are traditionally posed in a probabilistic framework, as was mentioned in the previous work sections (2.2.4, 2.2.3 and 2.3.2). The goal is usually to minimize the classifier probability error or some other performance criteria. Since the required knowledge of the associated probabilities is hard to obtain in practice, in recent years an effort has been done to develop for certain types of sources, agnostic methods where *universal* classification rules are applied, which are independent of the unknown underlying statistics and aiming to perform as well as the traditional methods [131, 40, 6, 116, 67, 22, 72, 98, 12, 4, 119, 48, 13, 50, 23, 24].

In this thesis, we aim to implement such *universal* classification rules (in the sense of being independent of the unknown sources models) and apply it to texts and ECG data sequences in the same way, thus disregarding any information about the type of source. In this manner, for example, we can make a text classification tool to be language or idiom independent. We will show that despite ignoring the a priori knowledge about the structure of the source, accurate classification can be achieved.

Basically, this research work started in 2004 and the first results were published a year later [22]. It follows and extends the same ideas of the information theoretic approaches proposed for sequence/text classification by Benedetto et al. [6] in 2002 and Li et al. [67] in 2004, which are a quite particular case of using symbols information based on lossless compression schemes. The key idea is to use the compression model acquired from one sequence to compress another sequence with off-the-shelf or adapted compression programs. If the two sequences are outputs from the same source, the resulting bit-wise size of the compressed file will be relatively low. Later on we discovered other authors that used similar compression-based approaches, such as [59], [72], [98] and [12].

The main idea behind this work is that to guess the correct class of the sequence **y**, given a collection $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k\}$ of representative sequences from the k classes, we may use the rule

$$\hat{w}(\mathbf{y}) = \arg\min_{\mathbf{x}_i \in \mathbf{X}} H(\mathbf{y}||\,\mathbf{x}_i)\,,$$

where $H(\mathbf{y}||\,\mathbf{x}_i)$ is some approximation to the *relative entropy* of the unknown sequence **y** with respect to other sequence $\mathbf{x}_i$ which is assumed to be a certain class model sequence. In this thesis, to obtain an approximation to the relative entropy, we propose the use of the Ziv-Merhav method as described in Section 3.6.

The use of compression-based methods in machine learning problems like the classification of several types of sequences has appeared in a variety of domains [6, 116, 72, 47, 13, 50]). While easy to apply, procedures like the one proposed by Kukushkina et al. [61], Benedetto et al. [6], and Li et al. [67], enable average computer users with access to off-the-shelf compression programs to easily perform classification. Several off-the-shelf compression algorithms have been used with this approach including RAR, LZW, PPM, GZIP, BZIP2, among others. These methods are effectively employed in applications on diverse data types with a basically parameter-free approach, which decreases the disadvantages of working with parameter-dependent algorithms. Clearly, in this work we found that, for certain applications, compression-based methods using *cross parsing* perform better than the well known *normalized compression distance* (NCD), a fact also stated recently by Helmer et al. [50].

Moreover, the main attraction of these compression-based methods for classification is that they avoid the problems of explicit feature extraction and selection, thus requiring virtually no preprocessing of the input sequence. For text classification, such methods do not require obtaining a representation of texts, like the bag-of-words, and the classification algorithm incorporates the quantification of textual properties.

Most such methods use compression models that describe the characteristics of the texts, usually based on repetitions of character sequences. In that sense, they can be considered character based, and thus have the potential to automatically capture contextual information and non-word features of a text, such as punctuation, word stems, and features spanning more than one word. Actually, some compression models, such as the Lempel-Ziv dictionaries, can be seen as an extension of the bag-of-words model.

So let us now stress some advantages of using the proposed approach:

- involves simple procedures;
- virtually need no preprocessing;
- enables the use of standard compression algorithms;
- can be effectively employed over diverse data types (*universal*).

However, compression-based classification methods have drawbacks; these algorithms may run quite slowly, thus are not suitable when speed is important, and are not equally effective for all application domains.

4.3 Dissimilarity-Based Classification

At the core of dissimilarity-based methods for classification is the computation of pairwise dissimilarities between the object (e.g., text) to be classified and a set of (or all) objects (e.g., texts) in the training set. Some of such methods are based on compression methods, which use in an unusual way off-the-shelf and other adapted compressors to obtain dissimilarity values. Examples of such techniques to obtain dissimilarity measures were briefly described in Chapter 3, namely the Normalized Compression Distance (NCD), the relative entropy estimated via the Ziv-Merhav Method (ZMM) and the Cross-Parsing Distance (CPD), which are all used in this work.

Of course, there are several ways to use dissimilarity values to define a classifier. In this thesis we consider the following two:

- using the dissimilarity values directly with a simple k-NN classifier, which we will refer as working with *'raw' dissimilarities*;

- using dissimilarity values as features and work in a *dissimilarity space*, since most other more sophisticated classifiers (e.g., SVM) work in a feature space naturally.

The simplest choice is arguably to use a k-NN classifier with *'raw' dissimilarities*; in this case, the object to be classified is simply assigned to the majority class in its k nearest neighbors in the adopted similarity measure (with some rule to break ties). We applied successfully this approach in our first works on text authorship attribution [22] (reproduced in Appendix A) and ECG biometrics [27] (reproduced in Appendix B).

The *dissimilarity space* approach, proposed by Pekalska et al. [84, 86], is more sophisticated and uses the dissimilarity values as features that characterize the object to be classified, based on which several different types of classifiers can be used, namely SVM, naïve Bayes or even k-NN in this *dissimilarity space*. This was the approach we adopted in our recent work on text sentiment analysis and authorship attribution [25] (reproduced in Appendix G), using as dissimilarity measures **ZMM** and **ZMMinc** respectively.

Formally, let us consider a training collection of objects (e.g., texts) $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, where each object belongs to some set $\mathcal{X}$ (e.g., the set of finite length strings of some finite alphabet Σ), and some dissimilarity measure between pairs of objects, $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In the *dissimilarity space* approach, each object (either in the training set or a new object to be classified after training) is represented by the vector of its dissimilarities with respect to the elements of $\mathbf{X}$ (or a subset thereof). That is, the training set in the so-called *dissimilarity space* becomes

$$\mathcal{D} = \{\mathbf{d}_1, ..., \mathbf{d}_n\},$$

where

$$\mathbf{d}_i = \begin{bmatrix} D(\mathbf{x}_i, \mathbf{x}_1) \\ \vdots \\ D(\mathbf{x}_i, \mathbf{x}_n) \end{bmatrix} \in \mathbb{R}^n.$$

In this paper, we propose to use a *dissimilarity space* approach, where object (feature) vector representations $\mathbf{d}_i$ are built by using dissimilarity/distance measures like the NCD, the Ziv-Merhav relative entropy estimates ZMM and ZMMinc, and the CPD, previously described in Chapter 3. Once in possession of a dissimilarity-based representation of a training set, any standard classification method can be used; in this paper, we report results based on k-NN and (linear) SVM classifiers.

For example, when using a k-NN classifier, given a new object $\mathbf{y}$, its dissimilarity vector $\mathbf{d}_y$ is built and distances in the *dissimilarity space* are computed (Euclidean distances, for example) between the new vector $\mathbf{d}_y$ and all the vectors in $\mathcal{D}$. Then, the new object $\mathbf{y}$ is classified in the most common class amongst its k nearest neighbors.

An important aspect of *dissimilarity space* approaches is that very few conditions are put of the dissimilarity measure; namely, it doesn't have to be a metric, it doesn't even need to be symmetric [84, 86].

4.4 Text Authorship Attribution

To handle the problem of text Authorship Attribution (AA) we use a *dissimilarity space* approach (for details see Appendixes F and G). Each text is mapped to a point in the feature space, where the features are the dissimilarity values between that text and all the class model texts; the class models are built upon string concatenation of all the training text samples (examples) from each class, i.e. author's book samples. The classification process is depicted in Figure 4.1.

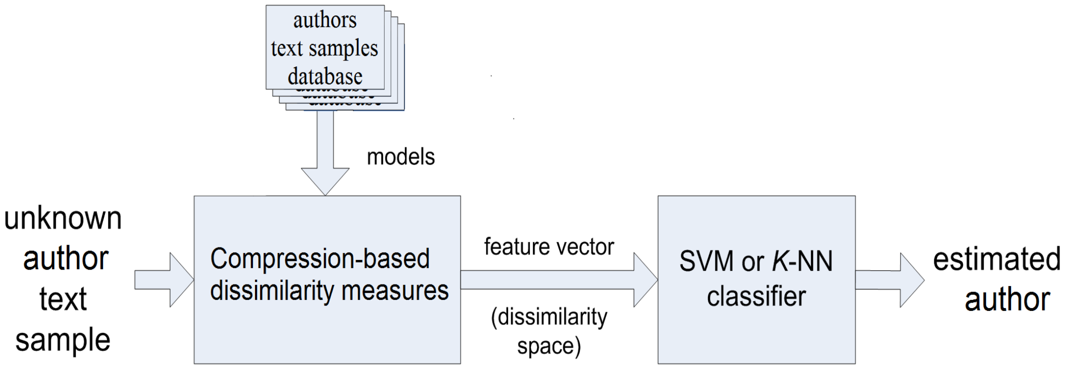


Figure 4.1: Proposed system block diagram for text authorship attribution.

For the AA experiments, we use three different corpora: (i) the Italian Corpus, introduced by Benedetto et al. [6], with 90 texts from 11 Italian authors[1] spanning the 13th to 20th century; (ii) the English Corpus, introduced by Ebrahimpour et al. [38], containing 168 short stories by seven undisputed English authors[2] from the late 19th century to the early 20th century, truncate to approximately the first 5,000 words, due to the differing lengths of the books; (iii) the Federalist Papers [76], where we consider only the 70 undisputed (out of 85) political essays published in 1788 by three American authors[3]. Due to the reduced number of texts samples in each corpus, we use leave-one-out cross-validation (LOO-CV) to assess the accuracy of the classifiers.

Experiments were done using the **NCD**, **ZMMinc** and **CPdist** as dissimilarity measures. In all the experiments, we do not use any text preprocessing. The k-NN and SVM classifiers (with linear kernel) used are implemented by the PRTools Matlab toolbox for pattern recognition [4] (version 4). The SVM penalty parameter C value was set to 1 or adjusted by

[1] Available at `http://www.liberliber.it`

[2] Available at `http://promo.net/pg`

[3] Available at `https://github.com/matthewberryman/author-detection/tree/master/Federalist%20Texts`

[4] Available at `http://www.prtools.org/index.html`

cross-validation. Reported results are in terms of the classification accuracy, expressed in percentage, with the accuracy assessed by LOO-CV.

Table 4.1 shows the accuracy results for each of the corpus. Our method **ZMMinc**, with optimized C, obtains an accuracy of 98.8% on the English Corpus (only fails 2 out of 168 texts), outperforming the methods of Ebrahimpour et al. [38]; on the two other corpora, the performance is approximately 4% below the baseline. Notice, however, that our results are obtained without any text preprocessing or any feature design/engineering, thus can be considered as highly competitive with those other methods.

Table 4.1: AA results: leave-one-out cross-validation accuracy percentages, using several dissimilarity measures with k-NN and SVM classifiers on 3 benchmark corpora.

		k-NN			**SVM**		
Corpus	**Baseline**	**NCD**	**CPdist**	**ZMMinc**	**NCD**	**CPdist**	**ZMMinc**
Italian	97.8	52.2	82.2	64.4	80.0	92.2	**94.4**
English	96.4	84.5	91.7	87.5	95.2	95.2	**98.8**
Federalist	97.1	82.9	90.0	84.3	62.2	81.4	**92.9**

4.5 Text Sentiment Analysis

We approach the problem of text Sentiment Analysis (SA) with classifiers also implemented in the *dissimilarity space*, where each text is mapped to a point in the feature space and the features are the dissimilarity values between that text and all the class model texts (see Appendixes E and G for details). The class models are also built upon string concatenation of the training text samples from each class, i.e. negative reviews text samples and positive reviews text samples. The classification process is depicted in Figure 4.2.

We conducted SA experiments on five well known datasets. Namely, we used the Movie Review Data[5] (more precisely the polarity dataset v2.0), introduced by Pang and Lee in 2004 [81], and the Multi-Domain Sentiment Dataset (version 2.0)[6], introduced by Blitzer et al. [10], which includes four datasets with Amazon reviews of four classes of products: Books, DVD, Electronics, and Kitchen. Each of the five datasets is labeled by humans and include 1,000 positive and 1,000 negative unprocessed reviews. We report 5-fold cross-validation (CV) accuracy estimates, following the same protocol of Xia et al. [127], where in each run 1600 examples are used to train and 400 examples to test.

[5] Available at `http://www.cs.cornell.edu/people/pabo/movie-review-data`

[6] Available at `http://www.cs.jhu.edu/~mdredze/datasets/sentiment`

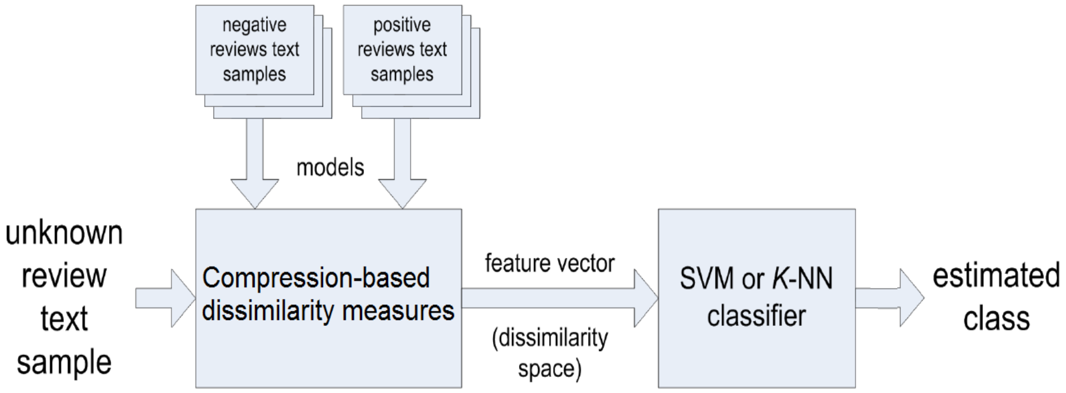

negative reviews text samples
positive reviews text samples
models
unknown review text sample
Compression-based dissimilarity measures
feature vector
(dissimilarity space)
SVM or K-NN classifier
estimated class

Figure 4.2: Block diagram of the proposed system for sentiment analysis.

In all the experiments we use the **NCD**, **ZMM** and **CPdist** as dissimilarity measures. We stress that, we do not use any text preprocessing. The k-NN and SVM classifiers (with linear kernel) used are from the same PRTools Matlab toolbox as used for AA experiments. The SVM penalty parameter (usually denoted by C) value was set to 1 or adjusted by CV. Reported results are in terms of the classification accuracy, expressed in percentage, with the accuracy assessed by CV.

Table 4.2 shows the 5-fold CV accuracy estimates of an SVM (with $C = 1$, except in the case denoted as **ZMMoptC**) working in the *dissimilarity space*. For comparison purposes, we also show the baseline and best results on the same datasets, described by Xia et al. [127]. Our method **ZMMoptC** achieves an accuracy of 82.41%, which is better than both baselines and is close to the best results reported by Xia et al. [127].

Table 4.2: SA results: 5-fold CV accuracy percentages, using several dissimilarity measures with SVM classifiers on 5 benchmark datasets. For comparison, we also show in the last four columns the results obtained by Xia et al. [127] over the same datasets, using the approaches POS-based (M1), part-of-speech information, and WR-based (M2), word relation features, plus the baselines assumed by those authors, respectively.

Dataset	**NCD**	**CPdist**	**ZMM**	**ZMMoptC**	**Baseline1**	**M1**	**Baseline2**	**M2**
Movies	84.85	80.45	84.60	85.80	84.75	86.80	86.45	87.70
Books	74.65	79.15	78.85	80.85	74.70	80.10	77.65	81.80
DVD	78.05	79.60	79.05	81.95	77.20	80.40	79.45	83.80
Elec	81.60	80.85	78.05	81.25	80.05	83.40	82.50	85.95
Kitchen	82.10	81.40	78.70	82.20	83.25	84.90	85.40	88.65
Average	**80.25**	**80.29**	**79.85**	**82.41**	**79.99**	**83.12**	**82.29**	**85.58**

4.6 ECG Biometrics

To built an ECG-based biometric system there are two different approaches in the literature concerning feature extraction: fiducial and non-fiducial. As mentioned in Section 2.3.1, non-fiducial methods base their decision directly on the waveform, without extracting intermediate features. In this work, following that same idea, we proposed for ECG classification a novel and simple non-fiducial method based on waveform comparison using information theoretic similarity measures. It uses an (optional) initial preprocessing step of single heartbeat waveforms segmentation and alignment by their R peaks. Moreover, it uses quantization to convert the ECG discrete-time analog signal values into a sequence of symbols (*i.e.* a string), followed by a *'raw' space* defined classifier based on string matching (for details see Appendixes B, C and D). Figure 4.3 depicts the proposed approach for ECG-based identification using a database with templates from N users.

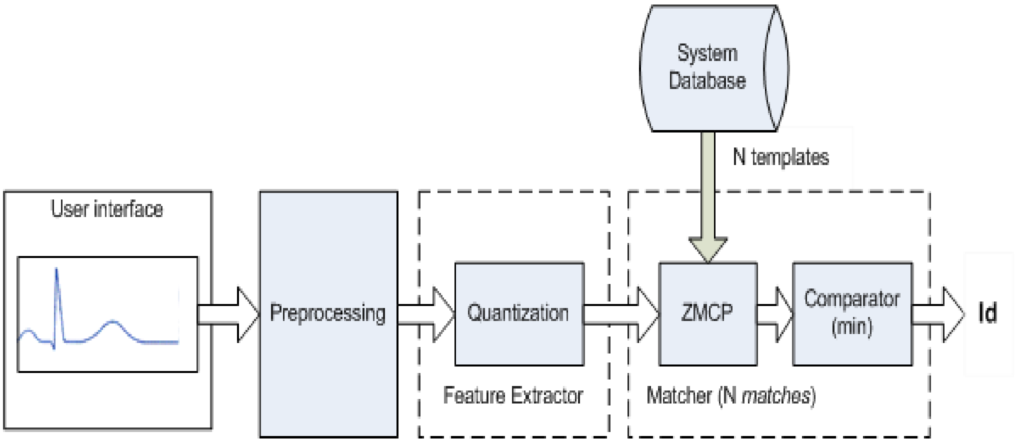


Figure 4.3: Block diagram of the proposed system for ECG-based identification given a database with templates from N users.

4.6.1 Quantization

The simplest approach to convert a set of single heartbeat waveforms into a set of strings is to apply N-bit uniform quantization, which produces sequences of symbols (strings) from an alphabet with 2^N symbols. Thus, a collection of heartbeat waveforms is transformed into a collection of strings. Subsequently, all the tools developed for text classification and string matching can be applied to ECG classification.

Quantization with less then 8 bits was considered in early experiments, but discarded because the resulting performance was lower than with 8-bit quantization. Despite the information loss due to the quantization process, our experimental results show that enough discriminative information is preserved. However, to improve the discriminative capability, we also tested non-uniform quantizers, such as Lloyd-Max quantizers, adapted and associated to each user. These quantizers use the well-known Lloyd-Max algorithm, which minimizes the MSQE (mean squared quantization error), and have been previously used in ECG compression for transmission purposes [92].

4.6.2 Classifier

The novel non-fiducial method proposed in this work is grounded in information theoretic text classification concepts and tools, namely the concept of cross complexity proposed by Cerra in 2009 [12]. Due to the reduced length of the strings (*i.e.* $\approx 220 - 256$ bytes) associated with each segmented single heartbeat waveform, we adopt a normalized cross parsing length as dissimilarity measure (in a similar fashion as the Cross Parsing Distance described in Section 3.5), defined by

$$C(\mathbf{z}|\mathbf{x}) = \tfrac{c(\mathbf{z}|\mathbf{x})}{|\,\mathbf{z}\,|},$$

which yields values in the range $[0, 1]$, that are compatible with the threshold levels defined for authentication purposes. In this definition, $|\,\mathbf{z}\,|$ is the (byte) length of the sequence $\mathbf{z}$. Notice that when the strings are very different, the estimated cross complexity will be close to $|\,\mathbf{z}\,|$, making $C(\mathbf{z}|\mathbf{x}) = 1$. For very similar strings, the estimated cross complexity will be low and thus C will be close to zero.

For identification, we use a 1-NN classifier working in a *'raw' space*, where a sample (string) $\mathbf{z}$ from an unknown subject is assigned to one of a set of K classes, given the subject (string) models $\mathbf{x}_k$ per class k, based on the computed value of the lowest dissimilarity measure, computed using the CP77 algorithm. In other words, the sample will be classified as belonging to the subject that leads to its shortest description. The classification rule is given by

$$\hat{k}(\mathbf{z}) = \arg \min_{k \in \{1,\ldots,K\}} C(\mathbf{z}|\mathbf{x}_k)\,.$$

For authentication, the classifier works in a *'raw' space* also using the CP77 algorithm, to determine the dissimilarity between a given unknown pattern $\mathbf{z}$ and the known template $\mathbf{x}_k$ in the database for whom the user claims to be, and compares it to a threshold. In our approach, we adopt a user-tuned threshold, previously established during the enrollment process for that particular user.

Experiments were done to assess the performance of the proposed methods, where we used two ECG databases: the HiMotion database [43], which has data from 26 subjects collected in an unrestrained scenario while performing a regular computer-based task, and the PTB diagnostics dataset [7], which has data from 51 healthy subjects collected in a clinical scenario. For comparison purposes, we also made experiments with a fiducial approch

[7] http://www.physionet.org/physiobank/database/ptbdb/

Table 4.3: Performance comparison of fiducial and non-fiducial approaches. The accuracy values (and standard deviation – std) refer to person identification, while the Equal error Rate (EER) values refer to authentication; Subj. is the number of subjects in the database, where the HiMotion ECG database and the PTB diagnostic ECG database were considered.

Database	Subj.	Approach	Identification accuracy (std)	Authentication EER (std)
HiMotion	26	Fiducial	99.57% (0.29%)	0.70% (0.15%)
HiMotion	26	Non-fiducial	99.94% (0.24%)	0.29% (0.95%)
PTB	51	Fiducial	99.85% (0.41%)	0.01% (0.02%)
PTB	51	Non-fiducial	99.39% (0.89%)	0.13% (0.42%)

over the same databases (see [28]for details). The adopted experiments strategy includes 50 (repetition) runs and LOO-CV (leave-on-out cross validation).

The achieved results are summarized in Table 4.3 for both fiducial and non-fiducial approaches over the considered databases, HiMotion ECG database and the PTB diagnostic ECG database.

Chapter 5

Conclusions and Future Work

In the emerging context of massive collections of online information, such as e-mail messages, music files, biometric data, product reviews and eBooks, for example, automatic data classification plays an important role, namely in the growing market of the handheld computers and smartphones applications. But learning how to classify such different types of data is very challenging. The problem of sequence classification has been widely studied for several communities, namely in the machine learning, data mining and information retrieval communities, with application domains like e-mail filtering, text authorship attribution or biometric recognition.

The primary goal of this thesis was to develop *universal* similarity measures (between sequences of symbols) and classification methods based on those measures, with application to text classification and ECG biometrics domains. In doing so, our contributions were:

- an original implementation of the *information theoretic dissimilarity measure* proposed by Ziv-Merhav, which is an empirical measure of the relative entropy between individual sequences that is based on self and cross parsing algorithms. We proposed an efficient cross-parsing algorithm based on the Lempel-Ziv sliding window algorithm and using optimized string matching data structures (suffix trees), expanding the empirical measure application domain to other type of sequences than finite-order Markovian sequences of the same size;

- a new way of using the dissimilarity measures for classification purposes. We used the dissimilarity measures as features to build a classifier in a *dissimilarity space*, where any classifier working in $\mathbb{R}^n$ can be used (i.e. k-NN or SVM). We expanded the state-of-the-art, by proposing a novel supervised classification method, which uses one of the different types available of compression-based measures to make *universal* sequence classification.

We have studied the problem of sequence classification ignoring a priori any information about the source model, specifically text and ECG classification ignoring either the linguistic structure of the texts and the P-QRS-T complexes structure of the ECG. Our classification methods do not require any specific preprocessing and handle sequences without distinction as in universal classification. Despite this fact, we have shown that our approach allows good results and it involves simple procedures when applying the classifier, as virtually no preprocessing (e.g., stop-word removal, word stemming, etc) is needed, nor feature extraction nor selection based on bag-of-words models. Also, it enables the use of adapted compression programs, that are based on universal lossless data compression algorithms, which use Lempel-Ziv dictionaries and can be effectively employed over diverse data types.

This research work started with the study and development of an implementation for the information theoretic dissimilarity measure proposed by Ziv-Merhav. Computational experiments showed that the developed implementation based on the LZ77 data compression algorithm works well as an empirical measure of the relative entropy between sequence as yields good estimates on synthetic Markov sequences. Moreover, this method was applied to a text classification problem (authorship attribution), outperforming a previously proposed approach over a corpus of 86 texts from Italian authors, while using no text preprocessing and a simple nearest neighbor classifier based on the developed measure. So, we conclude that it can be used as a tool for text classification.

The ECG-based biometrics problem was the next application domain. Due to the nature of the ECG single heartbeat waveform, namely its short length, and given that it is well known that the LZ77 algorithm is optimal as the sequence length $n \to \infty$, we proposed the use of a derived dissimilarity measure which relies on the cross parsing length of one (unknown) sequence given another (model) sequence. With this derived measure and a simple nearest neighbor classifier we proposed a new non-fiducial method to built biometrics systems.

To assess the absolute and relative performance of the new proposed method experiments were made using a fiducial approach also. For the fiducial approach, results have shown that, from a single mean waveform pattern, we were able to obtain a very good recognition accuracy, over both datasets (26 subjects from the HiMotion database and 51 subjects from the PTB database). Although, the non-fiducial approach results have shown that more than a single mean waveform pattern is needed as test sample to obtain similar results. But using 8 mean waveform patterns, we were able to obtain even better recognition accuracy over the HiMotion database, while obtaining as good recognition accuracy results over the PTB database. So, it's a matter of minimum information/samples that must be fed to the classifier in order to allow good performance. We conclude that the fiducial approach achieves very good performance with less data, *i.e.*, with a single mean heartbeat waveform used for test pattern. Nevertheless, the non-fiducial approach has the advantage of not requiring feature extraction, thus not relying critically on the detection of some fiducial points within the ECG signal. The final choice for one of the methods will depend on the system designer, given the application requirements and constraints.

More recently we come back to the text classification application domain, but now using dissimilarity measures as features to build a classifier in a *dissimilarity space*, where any classifier working in $\mathbb{R}^n$ can be used. Experiments were done using several information theoretic dissimilarity measures, namely normalized compression distance (NCD), relative entropy empirical estimate (ZMM) and cross parsing distance (CPD), for assessing the proposed methods performance on two classical text classification problems: sentiment (polarity) analysis (SA), which is a binary class problem, and authorship attribution (AA), which is a multi-class problem. We tested k-NN and SVM classifiers, where the best results were achieved with the latter.

Experimental results on SA reveal that the proposed methods approximate previous state-of-the-art techniques, despite being much simpler, in the sense that they do not require preprocessing and feature engineering. The SA experiments were done with 5 publicly available datasets with 1000 positive and 1000 negative reviews each. We conclude that the short length of the reviews could be an obstacle to the successful use of these compression-based dissimilarity measures. On the other hand, the experimental results on AA show that the proposed similarity measures and methods achieve good classification results and even set the state-of-the-art result in an authorship attribution problem, over an English Corpus where only fails 2 out of 168 texts.

Finally, according all experimental results in text classification, we conclude that the proposed information theoretic dissimilarity measure performs better than the Normalized Compression Distance (NCD) for the specific problems addressed.

In future work, we will aim at obtaining even better results, by using other kernels, other dissimilarity representations, and by exploiting the possibility of selecting a subset of objects with respect to which the dissimilarity representations are obtained. Some very important issues should be carefully studied: one is the influence of the sequence lengths on the efficiency of the proposed methods; other is the time consume assessment of the proposed methods and its comparison with the classical methods using BoW, feature extraction/selection and SVM for example. Furthermore, additional tests need to be done in order to assess the proposed approach in ECG-based biometrics without single heartbeat segmentation and even using classifiers working in the dissimilarity space. Other application domains must be considered and assessed, such as text categorization for example.

References

[1] CC Aggarwal and CX Zhai. A survey of text classification algorithms. In CC Aggarwal and CX Zhai, editors, *Mining text data*, chapter 6, pages 163–222. Springer, 2012.

[2] Foteini Agrafioti, Francis M. Bui, and Dimitrios Hatzinakos. Medical biometrics: the perils of ignoring time dependency. In *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, 2009. BTAS'09*, pages 1–6, 2009.

[3] Foteini Agrafioti, Jiexin Gao, and Dimitrios Hatzinakos. Heart Biometrics: Theory, Methods and Applications, Biometrics. In *Biometrics: Book 3*, pages 199–216. InTech, 2011.

[4] Alberto Apostolico and Fabio Cunial. Sequence Similarity by Gapped LZW. *2011 Data Compression Conference*, (2008217):343–352, March 2011.

[5] A. Azzini and S. Marrara. Impostor Users Discovery Using a Multimodal Biometric Continuous Authentication Fuzzy System. *Lecture Notes in Computer Science*, 5178:371–378, 2008.

[6] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. Language Trees and Zipping. *Physical Review Letters*, 88(4):048702, January 2002.

[7] C. H. Bennett, P. Gács, and Ming Li. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

[8] L. Biel, O. Pettersson, L. Philipson, and P. Wide. ECG analysis: a new approach in human identification. *IEEE Transactions on Instrumentation and Measurement*, 50(3):808–812, June 2001.

[9] Cristopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] John Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, 2007.

[11] Nikolaos V. Boulgouris, Konstantinos N. Plataniotis, and Evangelia Micheli-Tzanakou. *Biometrics: theory, methods, and applications*. Wiley-IEEE Press, 2009.

[12] Daniele Cerra and Mihai Datcu. Algorithmic Cross-Complexity and Relative Complexity. *2009 Data Compression Conference*, pages 342–351, March 2009.

[13] Daniele Cerra and Mihai Datcu. A fast compression-based similarity measure with applications to content-based image retrieval. *Journal of Visual Communication and Image Representation*, 23(2):293–302, February 2012.

[14] Daniele Cerra, Mihai Datcu, and Peter Reinartz. Authorship Analysis based on Data Compression. *to appear in Pattern Recognition Letters*, February 2014.

[15] Gregory J. Chaitin. On the Length of Programs for Computing Finite Binary Sequences: Statistical Considerations. *Journal of the ACM*, 13:547–569, 1969.

[16] ADC Chan and MM Hamdy. Wavelet distance measure for person identification using electrocardiograms. *Instrumentation and Measurement, IEEE Transactions on*, 57(2):248–253, 2008.

[17] Adrian D. C. Chan and M. M. Hamdy. Person identification using electrocardiograms. In *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 2006)*, number May, pages 1–4, 2006.

[18] Chuang-Chien Chiu, Chou-Min Chuang, and Chih-Yu Hsu. A Novel Personal Identity Verification Approach Using a Discrete Wavelet Transform of the ECG Signal. In *2008 International Conference on Multimedia and Ubiquitous Engineering (MUE 2008)*, pages 201–206. IEEE, 2008.

[19] E. Chung. *Pocketguide to ECG Diagnosis*. Blackwell Publishing, 2000.

[20] Rudi Cilibrasi and P. M. B. Vitányi. Clustering by compression. *Information Theory, IEEE Transactions on*, 51(4), 2005.

[21] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[22] David Pereira Coutinho and Mário A. T. Figueiredo. Information Theoretic Text Classification Using the Ziv-Merhav Method. In *Pattern Recognition and Image Analysis. Springer Berlin Heidelberg*, volume 1, pages 355–362, 2005.

[23] David Pereira Coutinho and Mário A. T. Figueiredo. An Information Theoretic Approach to Text Sentiment Analysis. In *Proceedings of 3rd International Conference on Pattern Recognition Applications and Methods (ICPRAM 2013)*, pages 577–580. SciTePress, 2013.

[24] David Pereira Coutinho and Mário A. T. Figueiredo. On Compression-Based Text Authorship Attribution. In *Proceedings of the 19th edition of the Portuguese Conference on Pattern Recognition (RECPAD 2013)*, number 4, 2013.

[25] David Pereira Coutinho and Mário A. T. Figueiredo. Text Classification Using Compression-based Dissimilarity Measures. *Submited to Pattern Recognition Letters*, 2013.

[26] David Pereira Coutinho, Ana L. N. Fred, and Mário A. T. Figueiredo. Personal Identification and Authentication based on One-lead ECG using Ziv-Merhav Cross Parsing. In *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems (PRIS 2010)*, pages 15–24, 2010.

[27] David Pereira Coutinho, Ana L.N. Fred, and Mário A.T. Figueiredo. One-Lead ECG-based Personal Identification Using Ziv-Merhav Cross Parsing. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010)*, pages 3858–3861. IEEE, August 2010.

[28] David Pereira Coutinho, Hugo Silva, Hugo Gamboa, Ana Fred, and Mário A. T. Figueiredo. Novel fiducial and non-fiducial approaches to electrocardiogram-based biometric systems. *IET Biometrics*, 2(2):64–75, 2013.

[29] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, I, 1967.

[30] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[31] J. Cunha, Bernardo Cunha, William Xavier, Nuno Ferreira, and A. Pereira. Vital-Jacket: A wearable wireless vital signs monitor for patients mobility. In *Proceedings of the Avantex Symposium*, 2007.

[32] I. G. Damousis, D. Tzovaras, and E. Bekiaris. Unobtrusive multimodal biometric authentication: The HUMABIO project concept. *EURASIP Journal on Advances in Signal Processing*, 2008:110, 2008.

[33] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(1-4):131–156, 1997.

[34] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, 29:103–130, 1997.

[35] Richard O. Duda, David G. Stork, and Peter E. Hart. *Pattern classification*. Wiley, New York, 2nd edition, 2000.

[36] Adnan Duric and Fei Song. Feature selection for sentiment analysis based on content and syntax models. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 96–103, 2011.

[37] I. K. Duskalov, Ivan A. Dotsinsky, and Ivailo I. Christov. Developments in ECG acquisition, preprocessing, parameter measurement, and recording. *IEEE Engineering in Medicine and Biology Magazine*, 17(2):50–58, 1998.

[38] Maryam Ebrahimpour, Tlis J. Putniš, Matthew J. Berryman, Andrew Allison, Brian W.-H. Ng, and Derek Abbott. Automated authorship attribution using advanced signal classification techniques. *PloS one*, 8(2):e54998, January 2013.

[39] Artur J. Ferreira and Mário A. T. Figueiredo. Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, 33(13):1794–1804, October 2012.

[40] Eibe Frank, Chang Chui, and I. H. Witten. Text categorization using compression models. 2000.

[41] H. Gamboa, H. Silva, and A. Fred. HiMotion Project. Technical report, 20070731, IT - Instituto de Telecomunicações, 2007.

[42] Hugo Gamboa. *Multi-Modal Behavioral Biometrics Based on HCI and Electrophysiology*. Phd thesis, Universidade Técnica de Lisboa, Instituto Superior Técnico, 2008.

[43] Hugo Gamboa, Hugo Silva, and Ana Fred. HiMotion: a new research resource for the study of behavior, cognition, and emotion. *Multimedia Tools and Applications*, July 2013.

[44] Jiexin Gao, Foteini Agrafioti, Hoda Mohammadzade, and Dimitrios Hatzinakos. ECG for blind identity verification in distributed systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1916–1919, 2011.

[45] Z. Geradts and A. Ruifrok. Extracting Forensic Evidence from Biometric Devices. *Proceedings of SPIE*, 5108:181–188, 2003.

[46] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation, Am. Heart Assoc.*, 101(23):e215–e220, 2000.

[47] Ramon De Graaff. *Authorship Attribution using Compression Distances*. Bachelor thesis, Leiden University, 2012.

[48] Ramon De Graaff and C. J. Veenman. Bootstrapped Authorship Attribution in Compression Space. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[49] Fredrik Gustafsson. Determining the initial states in forward-backward filtering. *IEEE Transactions on Signal Processing*, 44(4):988–992, 1996.

[50] Sven Helmer, Nikolaus Augsten, and Michael Böhlen. Measuring structural similarity of semistructured data based on information-theoretic approaches. *The VLDB Journal*, 21(5):677–702, February 2012.

[51] John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In *Artificial Intelligence: Methodology, Systems and Applications, ser. AIMSA'06*, pages 77–86, Berlin, 2006. Heidelberg: Springer-Verlag.

[52] Chih-Wei Hsu and Chih-Jen Lin. A Comparison of Methods for Multiclass Support Vector Machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[53] Steven A. Israel, John M. Irvine, Andrew Cheng, Mark D. Wiederhold, and Brenda K. Wiederhold. ECG to identify individuals. *Pattern Recognition*, 38(1):133–142, January 2005.

[54] A. Jain, P. Flynn, and A. Ross. *Handbook of Biometrics*. Springer, 2007.

[55] A. K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, January 2004.

[56] A. N. Jebaseeli and E. Kirubakaran. A Survey on Sentiment Analysis of(Product) Reviews. *International Journal of Computer Applications*, 47(11):36–39, 2012.

[57] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, 1998.

[58] Alistair Kennedy and Diana Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22(2):110–125, May 2006.

[59] Dmitry V. Khmelev and William J. Teahan. A repetition based measure for verification of text collections and for text categorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 104–110, New York, New York, USA, 2003. ACM Press.

[60] A. N. Kolmogorov. Three approaches to the quantitative definition ofinformation'. *Problems of information transmission*, 1(1):3–11, 1965.

[61] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2):172–184, 2001.

[62] U. Kunzmann, G. Wagner, J. Schöchlin, and A. Bolz. Parameter extraction of ECG signals in real-time. *Biomedizinische Technik/Biomedical Engineering*, 47(s1b):875–878, 2002.

[63] G. Kwang, R. Yap, T. Sim, and R. Ramnath. An Usability Study of Continuous Biometrics Authentication. *Lecture Notes in Computer Science*, 5558:828–837, 2009.

[64] Pablo Laguna, Roger G. Mark, A. Goldberg, and George B. Moody. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. In *Computers in Cardiology 1997*, pages 673–676. IEEE, 1997.

[65] N. J. Larsson. *Structures of string matching and data compression*. Phd thesis, Lund University, Sweden, 1999.

[66] Vladimir Leonov, Tom Torfs, Inge Doms, Refet Firat Yazicioglu, Ziyang Wang, Chris Van Hoof, and Ruud Vullers. Wireless body-powered electrocardiography shirt. In *Proceedings of the 3rd European Conference on Smart Systems Integration*, pages 307–314, 2009.

[67] M. Li, Xin Chen, and Xin Li. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.

[68] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 1997.

[69] R. Lourenço, P. Leite, A. Lourenço, H. Silva, A. Fred, and D. P. Coutinho. Experimental Apparatus for Finger ECG Biometrics. In *Proceedings of the International Conference on Biomedical Electronics and Devices (BIODEVICES 2012)*, pages 196–200, 2012.

[70] A. L. Maas, R. E. Daly, P. T. Pham, and Dan Huang. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics*, pages 142–150, 2011.

[71] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[72] Y. Marton, N. Wu, and L. Hellerstein. On compression-based text classification. *Advances in Information Retrieval*, pages 300–314, 2005.

[73] Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 301–311, 2005.

[74] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino. Impact of Artificial "Gummy" Fingers on Fingerprint Systems. *Proceedings of SPIE*, 4677:275–289, 2002.

[75] Liliana Medina and Ana Fred. Genetic Algorithm for Clustering Temporal Data-Application to the Detection of Stress from ECG Signals. In *Proceedings of 2nd International Conference on Agents and Artificial Intelligence (ICAART)*, pages 135–142, 2010.

[76] F. Mosteller and D. Wallace. *Inference and disputed authorship: The Federalist*. Addison-Wesley, 1964.

[77] F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[78] Mark Nelson and Jean-loup Gailly. *The Data Compression Book*. M&T Books, New York, 2nd editio edition, 1995.

[79] K. Niinuma and A. K. Jain. Continuous user authentication using temporal information. *Proceedings of SPIE*, 7667:76670L–76670L, 2010.

[80] Carla Oliveira and Ana L. N. Fred. ECG-based Authentication-Bayesian vs. Nearest Neighbour Classifiers. In *Proceedings of International Conference on Bio-inspired Systems and Signal Processing - Biosignals - INSTICC*, pages 163–168, 2009.

[81] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics*, page 271, 2004.

[82] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 1(2), 2008.

[83] Bo Pang, Lillian Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86, 2002.

[84] E. Pekalska, P. Paclik, and R. P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.

[85] Julien Penders, Bert Gyselinckx, Ruud Vullers, Olivier Rousseaux, Mladen Berekovic, Michael Nil, Chris Hoof, Julien Ryckaert, RefetFirat Yazicioglu, Paolo Fiorini, and Vladimir Leonov. Human++: Emerging technology for body area networks. In *VLSI-SoC: Research Trends in VLSI and Systems on Chip*, pages 377–397. Springer US, 2007.

[86] Elbieta Pkalska and Robert P. W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, 23(8):943–956, June 2002.

[87] Elbieta Pkalska, Robert P. W. Duin, and Pavel Paclík. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, February 2006.

[88] Konstantinos N. Plataniotis, Dimitrios Hatzinakos, and Jimmy K. M. Lee. ECG Biometric Recognition Without Fiducial Detection. *2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, pages 1–6, September 2006.

[89] A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani. Data compression and learning in time sequences analysis. *Physica D: Nonlinear Phenomena*, 180(1-2):92–107, June 2003.

[90] Alejandro Riera, Stephen Dunne, Ivan Cester, and Giulio Ruffini. Starfast: a wire-less wearable EEG/ECG biometric system based on the enobio sensor. In *Proceedings of the International Workshop on Wearable Micro and Nanosystems for Personalised Health*, 2008.

[91] Irina Rish. An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 3(22):41–46, 2001.

[92] M. Rodriguez, A. Ayala, S. Rodriguez, F. Rosa, and Mario Diaz-Gonzalez. Application of the MaxLloyd quantizer for ECG compression in diving mammals. *Computer Methods and Programs in Biomedicine*, 73(1):13–21, January 2004.

[93] Arun Abraham Ross, Karthik Nandakumar, and Anil K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, 2006.

[94] D. Salomon. *Data Compression. The complete reference*. Springer-Verlag New York, 3rd edition, 2004.

[95] David Salomon and Giovanni Motta. *Handbook of Data Compression*. Springer, 2010.

[96] Jorge Salvador Marques. *Reconhecimento de Padrões: Métodos Estatísticos e Neuronais*. IST Press, 1999.

[97] D. Sankoff and J. B. Kruskal. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Addison-Wesley, Reading, MA, 1983.

[98] D. Sculley and C.E. Brodley. Compression and Machine Learning: A New Perspective on Feature Space Vectors. *Data Compression Conference (DCC 2006)*, pages 332–332, 2006.

[99] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March 2002.

[100] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[101] A. Sharma and S. Dey. A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium. ACM*, pages 1–7, 2012.

[102] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[103] T. Shen. *Biometric Identity Verification Based on Electrocardiogram*. Phd thesis, University of Wisconsin, 2005.

[104] T. W. Shen, W. J. Tompkins, and Y. H. Hu. One-lead ECG for identity verification. In *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, pages 62–63, 2002.

[105] S. Shepherd. Continuous authentication by analysis of keyboard typing characteristics. In *Proceedings of the European Convention on Security and Detection*, pages 111–114, 1995.

[106] H. Silva, H. Gamboa, V. Viegas, and A. Fred. Wireless Physiologic Data Acquisition Platform. In *Proceedings of the 2005 Conference on Telecommunications*, 2005.

[107] H. Silva, A. Lourenço, R. Lourenç, P. Leite, D. P. Coutinho, and A. Fred. Study and evaluation of a single differential sensor design based on electro-textile electrodes for ECG biometrics applications. In *Proceedings of the IEEE Sensors Conference*, pages 1764–1767, 2011.

[108] Hugo Silva, Hugo Gamboa, and Ana Fred. Applicability of lead v2 ECG measurements in biometrics. In *Proceedings of Med-e-Tel*, 2007.

[109] Hugo Silva, Hugo Gamboa, and Ana Fred. One Lead ECG Based Personal Identification with Feature Subspace Ensembles. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2007)*, pages 770–783. Springer-Verlag Berlin, Heidelberg, 2007.

[110] H. H. So and K. L. Chan. Development of QRS detection method for real-time ambulatory cardiac monitor. In *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, volume 289, pages 289–292, 1997.

[111] R. J. Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1, March 1964):1–22, 1964.

[112] R. J. Solomonoff. A formal theory of inductive inference. Part II. *Information and control*, 7(2, June 1964):224–254, 1964.

[113] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[114] J. A. Storer and T. G. Szymanski. Data compression via textual substitution. *Journal of the ACM (JACM)*, 29(4):928–951, 1982.

[115] S. Suppappola and Y. Sun. A Comparison of Three QRS Detection Algorithms Using the AHA ECG Database. *IEEE Engineering in Medicine and Biology Society*, 13:586–587, 1991.

[116] W. J. Teahan and D. J. Harper. Using compression-based language models for text categorization. *Language Modeling for Information Retrieval*, 13:141–165, 2003.

[117] E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260, September 1995.

[118] G. Vinodhini and R. M. Chandrasekaran. Sentiment Analysis and Opinion Mining: A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6), 2012.

[119] P. M. B. Vitányi. Information distance: New developments. *arXiv preprint arXiv:1201.1221*, pages 1–4, 2012.

[120] Xuechuan Wang and Kuldip K. Paliwal. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern Recognition*, 36(10):2429–2439, October 2003.

[121] Yongjin Wang, Foteini Agrafioti, Dimitrios Hatzinakos, and Konstantinos N. Plataniotis. Analysis of Human Electrocardiogram for Biometric Recognition. *EURASIP Journal on Advances in Signal Processing*, 2008(1):19, 2008.

[122] Peter Weiner. Linear pattern matching algorithms. In *14th Annual Symposium on Found. of Computer Science (FOCS), Iowa City, Iowa*, pages 1–11, 1973.

[123] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using Appraisal Taxonomies for Sentiment Analysis. In *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*, 2005.

[124] John Wright, Yi Ma, Yangyu Tao, Zhouchen Lin, and Heung-Yeung Shum. Classification via Minimum Incremental Coding Length. *SIAM Journal on Imaging Sciences*, 2(2):367–395, January 2009.

[125] T. Wrublewski, Y. Sun, and J. Beyer. Real-time Early Detection of R Waves of the ECG Signals. *IEEE Engineering in Medicine and Biology Society*, 1:38–39, 1989.

[126] Gerd Wübbeler, Manuel Stavridis, Dieter Kreiseler, Ralf-Dieter Bousseljot, and Clemens Elster. Verification of humans using the electrocardiogram. *Pattern Recognition Letters*, 28(10):1172–1175, 2007.

[127] Rui Xia, Chengqing Zong, and Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6):1138–1152, March 2011.

[128] Ainur Yessenalina, Y. Yue, and C. Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics*, pages 1046–1056, 2010.

[129] J. Ziv and a. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977.

[130] J. Ziv and a. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, September 1978.

[131] J Ziv and N Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39(4):1270–1279, 1993.

Appendix A

Information Theoretic Text Classification Using the Ziv-Merhav Method

Paper A

Information Theoretic Text Classification Using the Ziv-Merhav Method

David Pereira Coutinho and Mário Figueiredo

In *Pattern Recognition and Image Analysis - IbPRIA 2005*,
LNCS 3523, pages 355–362, 2005

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.

A.1 Introduction

Defining a similarity measure between two finite sequences, without explicitly modeling 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.*, [6], 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 [1].

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 Loreto [1] 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 [11]. The ZM approach avoids the drawbacks of the method of Benedetto *et al* [1] which have been pointed out by Puglisi *et al* [5], 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 [1]. Our results show that ZM method outperforms the technique introduced in [1].

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.

A.2 Data Compression and Similarity Measures

A.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, 3], 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

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

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.*, [3]), that is

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

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 [1] and [11]. 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) [3].

A.2.2 Relationship Between Entropy and Lempel-Ziv Coding

Consider a sequence $\mathbf{x} = (x_1, x_2, ..., x_n)$ emitted by an unknown *l*th-order stationary Markovian source, defined over a finite alphabet. Suppose that one wishes to estimate the *n*th-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 [3], [10]): 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** can be approximated by

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

and it can be shown that it converges almost surely to $-(1/n)\log_2 p(x_1, x_2, ..., x_n)$, as $n \to \infty$ [11]. 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.

A.2.3 The Method of Benedetto, Caglioti and Loreto

Recently, Benedetto *et al* [1] 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, [1] 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

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

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) [1] defined a simplified "distance" function $d(A, B)$ between sequences,

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

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* [5] studied in detail what happens when a compression algorithm, such as LZ77 [9], 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.

A.2.4 Ziv-Merhav Empirical Divergence

The method proposed by Ziv and Merhav [11] for measuring relative entropy is also based on two Lempel-Ziv-type parsing algorithms:

- The incremental LZ parsing algorithm [10], 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 [11], 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

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

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 (A.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.

A.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 A.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.

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 A.1. At each step, the procedure is the same as with LZ77, except that the dictionary is not updated.

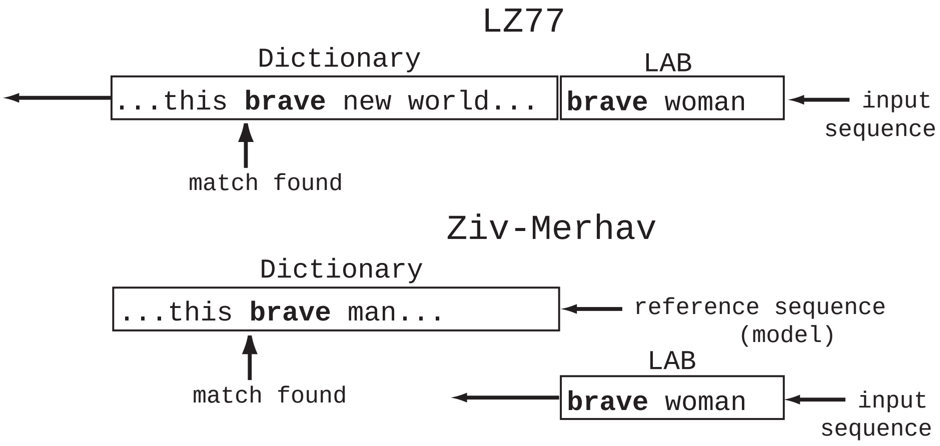


Figure A.1: The original LZ77 algorithm uses a sliding window over the input sequence to get the dictionary updated, whereas in the Ziv-Merhav cross parsing procedure the dictionary is static and only the *lookahead buffer* (LAB) slides over the input sequence.

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

A.4 Experiments

A.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 (A.2), while for the order-1 sources it is given by

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

Results for these experiments are shown in Figure A.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".

A.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* [1]. 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 [1], we apply both measures to this new corpus of Italian authors.

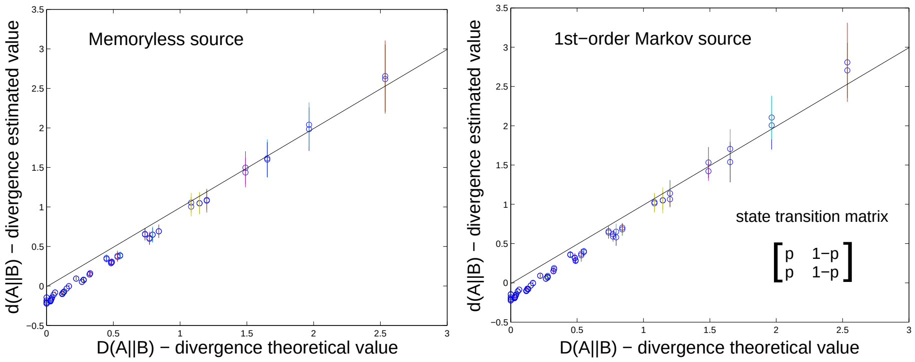


Figure A.2: Theoretical values versus Ziv-Merhav empirical divergence values, between two synthetic binary sequences of 10000 symbols length. Each circle is the sample mean value and the vertical segments are the sample standard deviation values, evaluated over 100 sequence pairs. For the 1st-order Markov 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.

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 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%.

A.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 [8].

References

[1] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. Language Trees and Zipping. *Physical Review Letters*, 88(4):048702, January 2002.

[2] C. H. Bennett, P. Gács, and Ming Li. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

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**

Table A.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).

[4] Mark Nelson and Jean-loup Gailly. *The Data Compression Book.* M&T Books, New York, 2nd editio edition, 1995.

[5] A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani. Data compression and learning in time sequences analysis. *Physica D: Nonlinear Phenomena*, 180(1-2):92–107, June 2003.

[6] D. Sankoff and J. B. Kruskal. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison.* Addison-Wesley, Reading, MA, 1983.

[7] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[8] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[9] J. Ziv and a. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977.

[10] J. Ziv and a. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, September 1978.

[11] J Ziv and N Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39(4):1270–1279, 1993.

Appendix B

One-Lead ECG-based Personal Identification Using Ziv-Merhav Cross Parsing

Paper B

One-Lead ECG-based Personal Identification Using Ziv-Merhav Cross Parsing

David Pereira Coutinho, Ana L. Fred and Mário Figueiredo

In *Proceedings of the 20th International Conference on Pattern Recognition - ICPR 2010*,
pages 3858–3861

Abstract

The advance of falsification technology increases security concerns and gives biometrics an important role in security solutions. The electrocardiogram (ECG) is an emerging biometric that does not need liveliness verification. There is strong evidence that ECG signals contain sufficient discriminative information to allow the identification of individuals from a large population. Most approaches rely on ECG data and the fiducia of different parts of the heartbeat waveform. However non-fiducial approaches have proved recently to be also effective, and have the advantage of not relying critically on the accurate extraction of fiducia data. In this paper, we propose a new non-fiducial ECG biometric identification method based on data compression techniques, namely the Ziv-Merhav cross parsing algorithm for symbol sequences (strings). Our method relies on a string similarity measure which can be seen as a compression-based approximation of the algorithmic cross complexity.We present results on real data, one-lead ECG, acquired during a concentration task, from 19 healthy individuals. Our approach achieves 100% subject recognition rate despite the existence of differentiated stress states.

B.1 Introduction

Biometrics deals with identification of individuals based on their physiological or behavioral characteristics [10] and plays an important role in security systems. Traditional methods of biometric identification, such as those using fingerprints or iris, provide accurate identification but lack robustness against falsification.

The electrocardiogram (ECG) is an emerging biometric tool exploiting a physiological feature that exists in all humans; there is a strong evidence that the ECG is sufficiently discriminative to identify individuals in a large population. The ECG has intrinsic liveliness verification, and allows personal identification and authentication, and detection of different stress or emotional states [11]. The ECG can also be used together with other biometric measures [13], as a complementary feature, for fusion in a multimodal system [2, Ch. 18] and for continuous verification where biological signatures are continuously monitored (easily done by using new signal acquisition technologies, such as the Vital Jacket [7]) in order to guarantee the identity of the operator throughout the whole process [8].

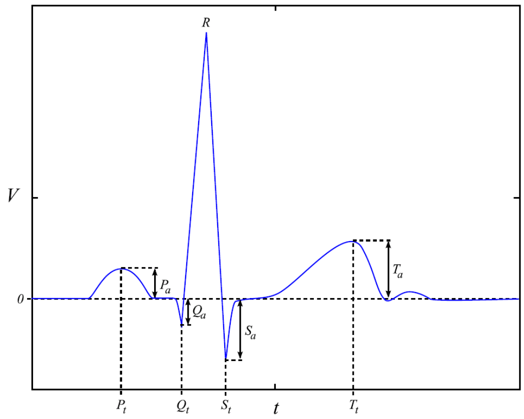


Figure B.1: Example of four latency times (features) measured from the P, QRS and T complexes of an ECG heartbeat for fiducial-based feature extraction.

A typical ECG signal of a normal heartbeat can be divided into 3 parts, as depicted in Figure B.1: the P wave (or P complex), which indicates the start and end of the atrial depolarization of the heart; the QRS complex, which corresponds to the ventricular depolarization; and, finally, the T wave (or T complex), which indicates the ventricular repolarization. It is known that the shape of these complexes differs from person to person, a fact which has stimulated the use of the ECG as a biometric [1].

In a broad sense, one can say there are two different approaches in the literature concerning feature extraction from ECG: fiducial [1] [15] [9] [16] and non-fiducial [4] [4]. Fiducial methods use points of interest within a single heartbeat waveform, such as local maxima or minima; these points are used as reference to allow the definition of latency times (features), as shown in Figure B.1. Several methods exist that extract different time and amplitude features, using these reference points. Non-fiducial techniques aim at extracting discriminative information from the ECG waveform without having to localize fiducial points. In this case, a global pattern from several heartbeat waveforms may be used as a feature. Some methods combine these two different approaches or are partially fiducial [17] (e.g., they use only the R peak as a reference for segmentation of the heartbeat waveforms). Table B.1 summarizes several approaches found in the literature; for more details on each method, see the corresponding publication.

Table B.1: Comparison of related work with our method. The accuracy (Accur.) values shown are the reported results for person identification.

Ref.	Feature	Method	Subjs.	Accur.
[1]	Fiducial	PCA	20	100%
[15]	Fiducial	Templ. matching+DBNN	20	100%
[9]	Fiducial	LDA	29	98 %
[16]	Fiducial	FSE	26	99.97%
[4]	Non-fiducial	Wavelet Distance	50	95%
[4]	Non-fiducial	Wavelet Distance	35	100%
[17]	Non-fiducial	AC/DCT+KNN	13	97.8%
Ours	Non-fiducial	Cross Parsing+MDL	19	100%

This paper introduces a new non-fiducial ECG biometric identification method that uses averaged single heartbeat waveforms and is based on data compression techniques, namely the Ziv-Merhav cross parsing algorithm for sequences of symbols, which derives from algorithmic cross complexity concept and its compression-based approximation. We present results on real data, using one-lead ECG acquisition during a concentration task. On a set of 19 healthy individuals, our method achieves 100% subject recognition rate despite the existence of differentiated stress states in the ECG signals [11].

The outline of the paper is as follows. In Section 2, we review the fundamental tools underlying our approach: Lempel-Ziv string parsing and compression; the Ziv-Merhav cross parsing algorithm. Section 3 presents the proposed classification method. Experimental results are presented in Section 4, while Section 5 concludes the paper.

B.2 The Lempel-Ziv and Ziv-Merhav Algorithms

The Lempel-Ziv (LZ) algorithm is a well-known tool for text compression [19] [20] [12] [14], which in recent years has also been used for classification purposes (see [5] and references therein). In particular, in [5], we have shown how the Ziv-Merhav (ZM) method for measuring relative entropy

[21] (which is based on Lempel-Ziv-type string parsing) achieves state-of-the-art performance in a specific text classification task. We will now briefly review these algorithms.

- The incremental LZ parsing algorithm [20], 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, $c(\mathbf{z}) = 6$.

- The ZM algorithm, a variant of the LZ parsing algorithm, is a sequential parsing of a sequence **z** with respect to another sequence **x** (cross parsing). Let $c(\mathbf{z}|\mathbf{x})$ denote the number of phrases in **z** with respect to **x**. For example, let **z** be as above and $\mathbf{x} = (10010100110)$; then, parsing **z** with respect to **x** yields $(011, 110, 00110)$, that is $c(\mathbf{z}|\mathbf{x}) = 3$.

Roughly speaking, we can see $c(\mathbf{z})$ as a measure of the complexity of the sequence **z**, while $c(\mathbf{z}|\mathbf{x})$, the description length obtained when coding **z** using a model for **x** (cross parsing), can be seen as an estimate of the cross complexity [3]. It is expectable that the cross complexity is low when the two sequences are very similar; this is the key idea behind the use of ZM cross parsing in classification [5], which in this paper will be adopted for ECG-based personal identification.

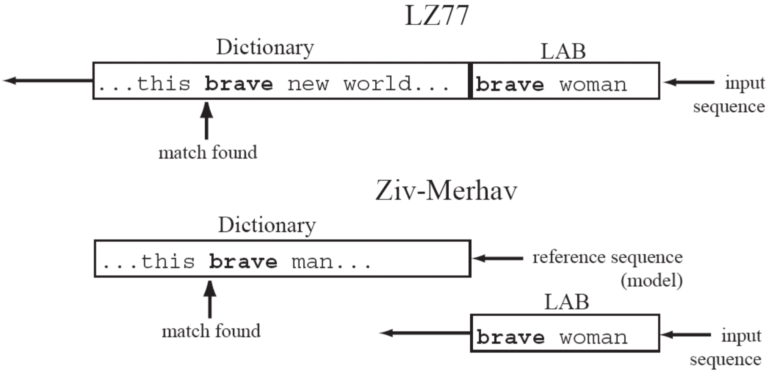


Figure B.2: The original LZ77 algorithm uses a sliding window over the input sequence to update the dictionary; in our implementation of ZM cross parsing, the dictionary is static and only the lookahead buffer (LAB) slides over the input sequence.

An implementation of the ZM cross parsing algorithm was proposed in [5], based on a modified LZ77 [19] algorithm, where the dictionary is static and only the lookahead buffer slides over the input sequence, as shown in Figure B.2 (for more details, see [5]). This very same implementation, using a 64 Kbyte dictionary and a 256 byte look ahead buffer, was used in the experiments reported below.

B.3 Proposed Classification Method

To use ZM-based tools for classification, a necessary first step is the conversion of the ECG (discrete-time analog) signal into a sequence of symbols. In this paper, we propose a very simple approach based on quantization. Assuming we are given a set of single heartbeat waveforms (resulting from a segmentation preprocessing stage), we simply apply 8-bit (256 levels) uniform quantization, thus obtaining a sequence of symbols (from a 256 symbols alphabet) from each single heartbeat.

Consider a collection of training samples partitioned into K classes (the set of subjects to be identified): $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2,..., \mathcal{X}_K\}$. For each subject/class k, $\mathcal{X}_k$ contains n strings obtained from the same number of heartbeats using the quantization procedure described in the previous paragraph. A string $\mathbf{x}_k$ is formed by concatenating the n training strings of subject k; string $\mathbf{x}_k$ is, in some sense, a "model" representing the shape of the heartbeats of subject k.

Given a test sample $\mathbf{z}$ (containing the string representing m heartbeats) obtained from an unknown subject (assumed to be one from which the training set was obtained), its identity is estimated as follows:

$$\hat{k}(\mathbf{z}) = \arg \min_{k \in \{1,...,K\}} c(\mathbf{z}|\mathbf{x}_k),$$

where $c(\mathbf{z}|\mathbf{x}_k)$ is computed by the ZM cross parsing algorithm, as described in Section B.2. In other words, the test sample is classified as belonging to the subject that leads to its shortest description. Although using different tools, this approach is related in spirit with the *minimum incremental coding length* (MICL) approach [18].

B.4 Experiments

B.4.1 Data collection

The ECG waveform dataset used was acquired using one lead, in the context of the Himotion project. The dataset contains ECG recordings from 19 subjects acquired during a concentration task on a computer, designed for an average completion time of 10 minutes. All the acquired ECG signals were normalized and band-pass filtered (2–30Hz) in order to remove noise. Each heartbeat waveform was sequentially segmented from the full recording and then all the obtained waveforms were aligned by their R peaks. From the resulting collection of ECG heartbeat waveforms, the mean wave for groups of 10 consecutive waveforms (without overlap) was computed. Each of these mean waveforms is what we call a single heartbeat in Section B.3. Notice that an intra-class study [11] with the dataset, in the context of the exploration of electrophysiological signals for emotional states detection, showed the existence of differentiated states in the data that represent the ECG signal of a subject.

B.4.2 Experimental Results

The reported results are averages over 30 runs. In each run, we partition the set of heartbeats of each subject into two mutually exclusive subsets: one of these subsets is used to form the training data set $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_K\}$, while the other is used to build the test waveforms. We consider several values for n (the number of "model" strings) as well as for m (the number of test waveforms).

The results of this experiment, which are depicted in Figure B.3, show that the proposed method achieves 100% accuracy for $m = 12$ and $n = 13$ or $n = 20$. This is better than the results reported in [16] over the same dataset. The approaches in [1], [15], [4], were not tested on the same dataset, so the results are not directly comparable. Notice that using only $m = 5$ waveforms for the test patterns, we already reach an accuracy around 99.5%. As expected, the accuracy increases both with n and m.

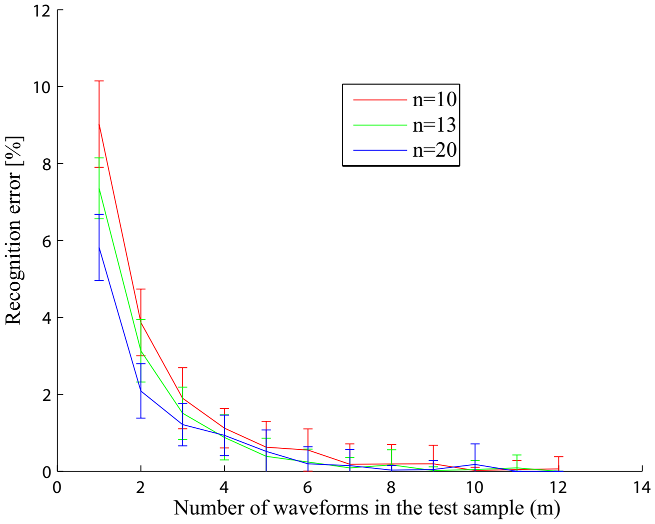


Figure B.3: Mean recognition error and standard deviation intervals for subject identification when considering a variable number of waveforms as test samples.

B.5 Summary and Conclusions

We have presented a method for personal identification from one-lead ECG signals which involves no explicit feature extraction other than 8 bit uniform quantization of the waveforms. Furthermore, after acquisition and preprocessing, the enrollment process relies only on the concatenation of quantized waveforms to build the models. The classifier is based on the Ziv-Merhav cross parsing algorithm [21], an estimator of the algorithmic cross complexity [3], which is used to measure similarity between the test waveform and the waveforms present in the training dataset. Experiments carried out on a set of 19 subjects showed that our method achieves 100% accuracy, outperforming a previously proposed approach [16] over the same dataset. Notice that this accuracy is achieved despite the existence of differentiated stress states in the dataset samples [11]. Although further experiments, on other datasets, are needed to assess the relative performance of the proposed method, with respect to other state-of-the-art techniques, these results demonstrated the validity of our approach as a tool for personal identification and of the ECG signal as a viable biometric. Moreover, this biometric trait can be easily acquired, or even continuously monitored using new acquisition technologies such as the Vital Jacket [7], and included in a multimodal system. Current work include further evaluation of our method when used for authentication purposes, with ROC curve design for false acceptance rate (FAR) and false rejection rate (FRR) analysis [6].

Acknowledgments

We acknowledge the following financial support: the FET programme, within the EU FP7, under the SIMBAD project (contract 213250); Fundação para a Ciência e Tecnologia (FCT), under grants PTDC/EEA-TEL/72572/2006 and QREN 3475.

References

[1] L. Biel, O. Pettersson, L. Philipson, and P. Wide. ECG analysis: a new approach in human identification. *IEEE Transactions on Instrumentation and Measurement*, 50(3):808–812, June 2001.

[2] Nikolaos V. Boulgouris, Konstantinos N. Plataniotis, and Evangelia Micheli-Tzanakou. *Biometrics: theory, methods, and applications*. Wiley-IEEE Press, 2009.

[3] Daniele Cerra and Mihai Datcu. Algorithmic Cross-Complexity and Relative Complexity. *2009 Data Compression Conference*, pages 342–351, March 2009.

[4] ADC Chan and MM Hamdy. Wavelet distance measure for person identification using electrocardiograms. *Instrumentation and Measurement, IEEE Transactions on*, 57(2):248–253, 2008.

[5] David Pereira Coutinho and Mário A. T. Figueiredo. Information Theoretic Text Classification Using the Ziv-Merhav Method. In *Pattern Recognition and Image Analysis. Springer Berlin Heidelberg*, volume 1, pages 355–362, 2005.

[6] David Pereira Coutinho, Ana L. N. Fred, and Mário A. T. Figueiredo. Personal Identification and Authentication based on One-lead ECG using Ziv-Merhav Cross Parsing. In *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems (PRIS 2010)*, pages 15–24, 2010.

[7] J. Cunha, Bernardo Cunha, William Xavier, Nuno Ferreira, and A. Pereira. Vital-Jacket: A wearable wireless vital signs monitor for patients mobility. In *Proceedings of the Avantex Symposium*, 2007.

[8] I. G. Damousis, D. Tzovaras, and E. Bekiaris. Unobtrusive multimodal biometric authentication: The HUMABIO project concept. *EURASIP Journal on Advances in Signal Processing*, 2008:110, 2008.

[9] Steven A. Israel, John M. Irvine, Andrew Cheng, Mark D. Wiederhold, and Brenda K. Wiederhold. ECG to identify individuals. *Pattern Recognition*, 38(1):133–142, January 2005.

[10] A. K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, January 2004.

[11] Liliana Medina and Ana Fred. Genetic Algorithm for Clustering Temporal Data-Application to the Detection of Stress from ECG Signals. In *Proceedings of 2nd International Conference on Agents and Artificial Intelligence (ICAART)*, pages 135–142, 2010.

[12] Mark Nelson and Jean-loup Gailly. *The Data Compression Book*. M&T Books, New York, 2nd editio edition, 1995.

[13] Arun Abraham Ross, Karthik Nandakumar, and Anil K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, 2006.

[14] D. Salomon. *Data Compression. The complete reference*. Springer-Verlag New York, 3rd edition, 2004.

[15] T. W. Shen, W. J. Tompkins, and Y. H. Hu. One-lead ECG for identity verification. In *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, pages 62–63, 2002.

[16] Hugo Silva, Hugo Gamboa, and Ana Fred. One Lead ECG Based Personal Identification with Feature Subspace Ensembles. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2007)*, pages 770–783. Springer-Verlag Berlin, Heidelberg, 2007.

[17] Yongjin Wang, Foteini Agrafioti, Dimitrios Hatzinakos, and Konstantinos N. Plataniotis. Analysis of Human Electrocardiogram for Biometric Recognition. *EURASIP Journal on Advances in Signal Processing*, 2008(1):19, 2008.

[18] John Wright, Yi Ma, Yangyu Tao, Zhouchen Lin, and Heung-Yeung Shum. Classification via Minimum Incremental Coding Length. *SIAM Journal on Imaging Sciences*, 2(2):367–395, January 2009.

[19] J. Ziv and a. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977.

[20] J. Ziv and a. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, September 1978.

[21] J Ziv and N Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39(4):1270–1279, 1993.

Appendix C

Personal Identification and Authentication based on One-lead ECG using Ziv-Merhav Cross Parsing

Paper C

Personal Identification and Authentication based on One-lead ECG using Ziv-Merhav Cross Parsing

Coutinho, D. P., Fred, A. L., and a.T. Figueiredo, M.

In *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems - PRIS 2010*, pages 15-24, 2010

Abstract

In this paper, we propose a new data compression based ECG biometric method for personal identification and authentication. The ECG is an emerging biometric that does not need liveliness verification. There is strong evidence that ECG signals contain sufficient discriminative information to allow the identification of individuals from a large population. Most approaches rely on ECG data and the fiducia of different parts of the heartbeat waveform. However non-fiducial approaches have proved recently to be also effective, and have the advantage of not relying critically on the accurate extraction of fiducia. We propose a non-fiducial method based on the Ziv-Merhav cross parsing algorithm for symbol sequences (strings). Our method uses a string similarity measure obtained with a data compression algorithm. We present results on real data, one-lead ECG, acquired during a concentration task, from 19 healthy individuals, on which our approach achieves 100% subject identification rate and an average equal error rate of 1.1% on the authentication task.

C.1 Introduction

Biometrics deals with identification of individuals based on their physiological or behavioral characteristics [11]. Traditional methods of biometric identification, include those based on physiological characteristics like fingerprints or iris, and those based on behavioral characteristics like signature

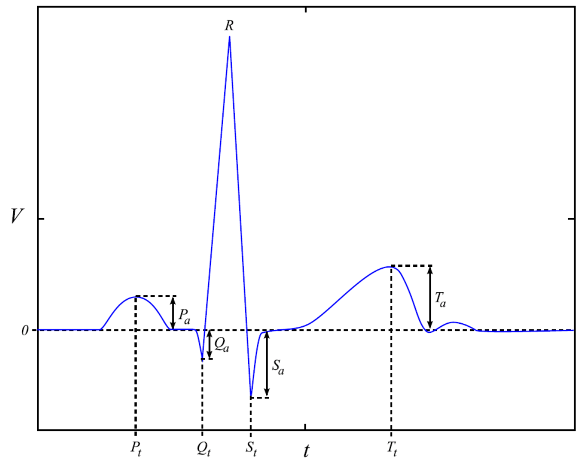


Figure C.1: Example of four latency times (features) measured from the P, QRS and T complexes of an ECG heartbeat for fiducial-based feature extraction.

or speech. Although some technologies have gained more acceptance than others, the field of biometrics for access control plays an important role in the security at airports, industry and corporate workplaces, for example. But some technologies lack robustness against falsification. Some may be based on such characteristics that for a group of people is difficult to acquire or even that characteristics is missing.

The electrocardiogram (ECG) is an emerging biometric measure which exploits a physiological feature that exists on every human and there is a strong evidence that the ECG is sufficiently discriminative to identify individuals from a large population. The ECG feature allows liveliness detection (intrinsic), personal identification and authentication, and different stress or emotion states detection [14]. The ECG is a behavioral biometric trait that can be used with other biometric measures [17], as a complementary feature, for fusion in a multimodal physiological authentication system [2, Ch. 18] and for continuous authentication where biological signatures are continuously monitored (easily done by using new signal acquisition technologies like the Vital Jacket [7], [13]) in order to guarantee the identity of the operator throughout the whole process [8].

A typical ECG signal of a normal heartbeat can be divided into 3 parts, as depicted in Fig. C.1: the P wave (or P complex), which indicates the start and end of the atrial depolarization of the heart; the QRS complex, which corresponds to the ventricular depolarization; and, finally, the T wave (or T complex), which indicates the ventricular repolarization. It is known that the shape of these complexes differs from person to person, a fact which has stimulated the use of the ECG as a biometric [1].

In a broad sense, one can say there are two different approaches in the literature concerning feature extraction from ECG: fiducial [1], [19], [10], [20], and non-fiducial [4], [5]. Fiducial methods use points of interest within a single heartbeat waveform, such as local maxima or minima; these points are used as reference to allow the definition of latency times, as shown in Fig. C.1. Several methods exist that extract different time and amplitude features, using these reference points. Non-fiducial techniques aim at extracting discriminative information from the ECG waveform without having to localize fiducial points. In this case, a global pattern from several heartbeat waveforms may be used as a feature. Some methods combine these two different approaches or are partially fiducial [22] (e.g., they use only the R peak as a reference for segmentation of the heartbeat waveforms).

Biel et al. [1] pioneered the use of the ECG as a biometric for personal identification. They used a 12-lead ECG but ended up concluding that one lead was enough because 12-lead ECG systems require meticulous placement of the electrodes on each person, which is not practical. Using a proprietary equipment from SIEMENS, 30 fiducial features were extracted; a feature selection algorithm allowed concluding that the best results were with 10 features. Classification was based on the principal component analysis (PCA) of each class. The purpose was to identify 20 subjects at rest and they achieved an accuracy of 100%.

Recent studies have shown that non-fiducial approaches also allow successful personal identification using the ECG heartbeat signal.

Chiu et al. [5], using a one-lead ECG, introduced a system based on a 3-step feature extraction method. It uses QRS complex detection (with the o and Chanmethod [21]) and waveform alignment in the time domain; the features extracted are based on the discrete wavelet transform. A nearest neighbor classifier based on the Euclidean distance between pairs of feature vectors is used. The purpose was to identify 35 subjects (no activity specified) from the QT database [12]. The results obtained were: 100% of accuracy on person identification and 0.83% FAR (false acceptance rate) and 0.86% FRR (false rejection rate) for authentication.

This paper introduces a new non-fiducial ECG-biometric method that uses averaged single heartbeat waveforms and is based on data compression techniques, namely the Ziv-Merhav cross parsing (ZMCP) algorithm for sequences of symbols. We present results on real data, using one-lead ECG acquisition during a concentration task. Notice that a study [14] with the dataset showed the existence of differentiated states in the data representing the ECG signal of a subject due to detectable changes along the time in the acquired signal. On a set of 19 healthy individuals, our method achieves 100% subject identification (recognition) rate and an average equal error rate of 1.1% on the authentication (verification) task.

The outline of the paper is as follows. In Section 2, we review the fundamental tools underlying our approach: Lempel-Ziv string parsing and compression; the Ziv-Merhav cross parsing algorithm. Section 3 presents the proposed classification method. Experimental results are presented in Section 4, while Section 5 concludes the paper.

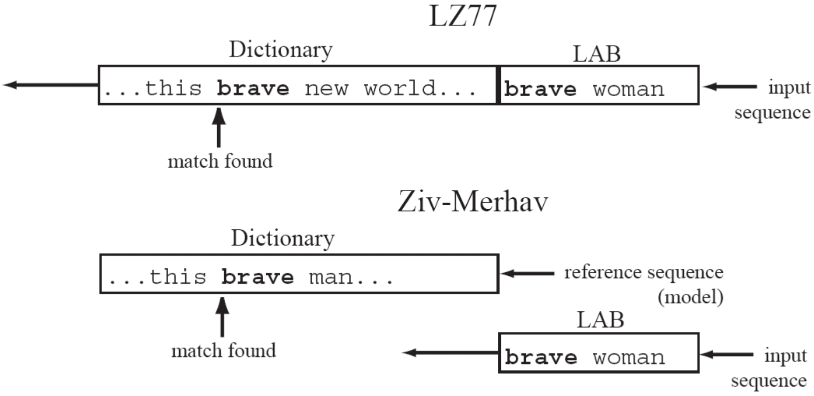


Figure C.2: The original LZ77 algorithm uses a sliding window over the input sequence to update the dictionary; in our implementation of ZM cross parsing algorithm, the dictionary is static and only the lookahead buffer (LAB) slides over the input sequence.

C.2 The Lempel-Ziv and Ziv-Merhav Algorithms

The Lempel-Ziv (LZ) algorithm is a well-known tool for text compression [24], [25], [15], [18], which in recent years has also been used for classification purposes (see [6] and references therein). In particular, in [6], we have shown how the Ziv-Merhav (ZM) method for measuring relative entropy [26] (which is based on Lempel-Ziv-type string parsing) achieves state-of-the-art performance in a specific text classification task. We will now briefly review these algorithms.

- The incremental LZ parsing algorithm [25], 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, $c(\mathbf{z}) = 6$.

- The ZM (cross parsing) algorithm, a variant of the LZ parsing algorithm, 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 $\mathbf{z}$ be as above and $\mathbf{x} = (10010100110)$; then, parsing $\mathbf{z}$ with respect to $\mathbf{x}$ yields $(011, 110, 00110)$, that is $c(\mathbf{z}|\mathbf{x}) = 3$.

Roughly speaking, we can see $c(\mathbf{z})$ as a measure of the complexity of the sequence $\mathbf{z}$, while $c(\mathbf{z}|\mathbf{x})$, the code-length obtained when coding $\mathbf{z}$ using a model for $\mathbf{x}$ (cross parsing), can be seen as an estimate of the cross complexity [3]. It is expectable that the cross complexity is low when the two sequences are very similar; this is the key idea behind the use of ZM cross parsing in classification [6], which in this paper will be adopted for ECG-based personal identification and authentication.

An implementation of the ZM cross parsing algorithm as a component of a ZM method for relative entropy estimation was proposed in [6], based on a modified LZ77 [24] algorithm, where the dictionary is static and only the lookahead buffer slides over the input sequence, as shown in Fig. C.2

(for more details, see [6]). This very same implementation of the cross parsing, using a 64 Kbyte dictionary and a 256 byte look ahead buffer, was used in the experiments reported below.

C.3 Proposed Methods

To use ZM-based tools for identification or authentication, a necessary first step is the conversion of the ECG (discrete-time analog) signal into a sequence of symbols (text). In this paper, we propose a very simple approach based on quantization. Assuming we are given a set of single heartbeat waveforms (resulting from a segmentation preprocessing stage), we simply apply 8-bit (256 levels) uniform quantization, thus obtaining a sequence of symbols (from a 256 symbols alphabet) from each single heartbeat.

Quantizers with fewer bits were considered in early experiments but discarded because they didn't perform as well as the 8-bit quantizer. Higher values were not considered for sake of system implementation simplicity and because of the good performance obtained with 8 bits.

Consider a collection of training samples partitioned into K classes (the set of subjects to be identified): $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_K\}$. For each subject/class k, $\mathcal{X}_k$ contains n strings obtained from the same number of heartbeats using the quantization procedure described in the previous paragraph. A string $\mathbf{x}_k$ is formed by concatenating the n training strings of subject k; string $\mathbf{x}_k$ is, in some sense, a "model" representing the shape of the heartbeats of subject k.

C.3.1 Identification

Given a test sample $\mathbf{z}$ (containing the string representing m heartbeats) obtained from an unknown subject (assumed to be one from which the training set was obtained), its identity is estimated as follows:

$$\hat{k}(\mathbf{z}) = \arg \min_{k \in \{1,\ldots,K\}} c(\mathbf{z}|\mathbf{x}_k),$$

where $c(\mathbf{z}|\mathbf{x}_k)$ is computed by the ZM cross parsing (ZMCP) algorithm, as described in Section C.2. In other words, the test sample is classified as belonging to the subject that leads to its shortest description. Although using different tools, this approach is related in spirit with the *minimum incremental coding length* (MICL) approach [23].

C.3.2 Authentication

The authentication (verification) procedure depends on a threshold level, which depends itself from the range of values of $c(\mathbf{z}|\mathbf{x}_k)$. In order to limit its variation to a predefined set of values, normalization is used. Since in the worst case the description length, resulting from the ZMCP algorithm for the test sample $\mathbf{z}$, is the length of $\mathbf{z}$, the normalized description length $c_n(\mathbf{z}|\mathbf{x}_k)$ is defined as follows:

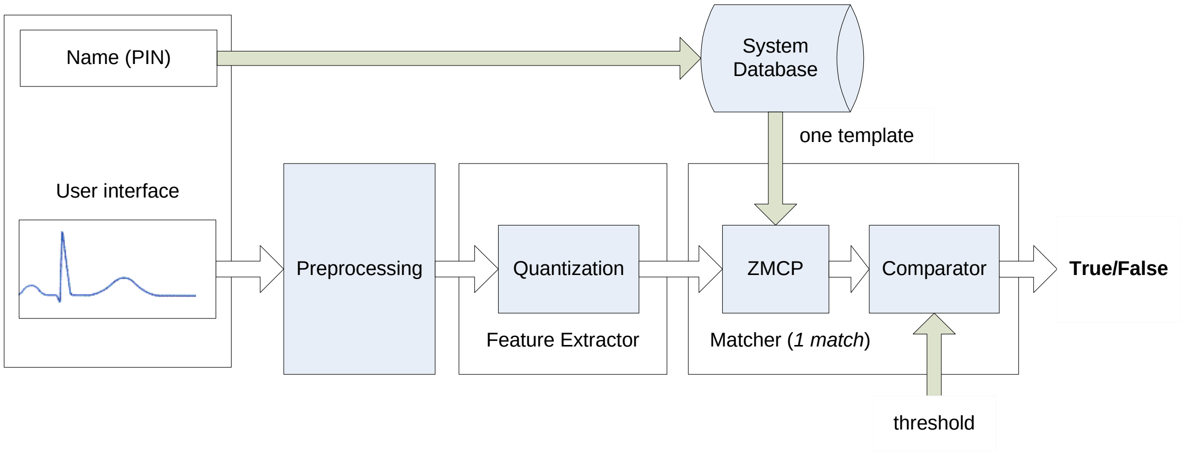


Figure C.3: Block diagram of the implemented system, for the authentication task, is shown using the five main modules of a biometric system, i.e., sensor, preprocessing, feature extraction, matcher and system database.

$$c_n(\mathbf{z}|\mathbf{x}_k) = \frac{c(\mathbf{z}|\mathbf{x}_k)}{len(\mathbf{z})},$$

where $len(\mathbf{z})$ is the number of bytes in test sample $\mathbf{z}$. Notice that $c_n(\mathbf{z}|\mathbf{x}_k) \in [0, 1]$.
Test sample verification is made by comparing the value of $c_n(\mathbf{z}|\mathbf{x}_k)$ when using the claimed identity model with a threshold value $\in [0, 1]$, previously set according to a selected error rate, false acceptance rate (FAR), or false rejection rate (FRR). It decides for genuine when the comparison result is less or equal to the selected threshold level.

C.4 Experiments

The architecture of the proposed ECG-based biometric system for person identification and authentication follows the same model proposed by Jain et al in [11]. Fig. C.3 shows the block diagram of the implemented system for the authentication task.

C.4.1 Data collection

The ECG waveform dataset used was acquired using one lead, in the context of the Himotion project.[1] The dataset contains ECG recordings from 19 subjects acquired during a concentration task on a computer, designed for an average completion time of 10 minutes. All the acquired ECG signals were normalized and band-pass filtered (2–30Hz) in order to remove noise. Each heartbeat waveform was sequentially segmented from the full recording and then all the obtained waveforms were aligned by their R peaks. From the resulting collection of ECG heartbeat waveforms, the mean wave for groups

[1] `https://www.it.pt/auto_temp_web_page_preview.asp?id=305`

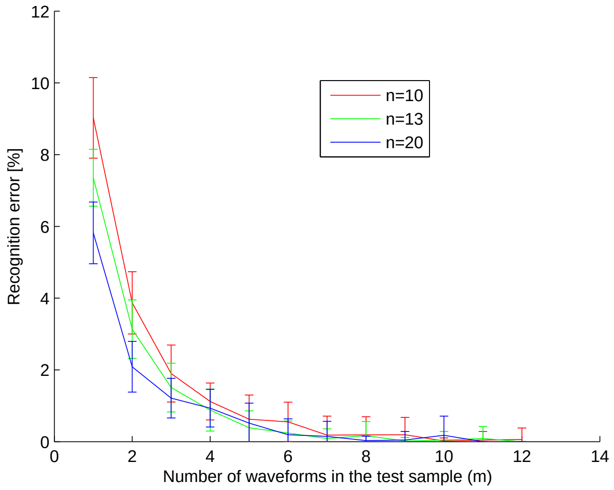


Figure C.4: Mean recognition error and standard deviation intervals for subject identification when considering a variable number of waveforms as test samples.

of 10 consecutive waveforms (without overlap) was computed. Each of these mean waveforms is what we call a single heartbeat in Section C.3.

An intra-class study [14] with the dataset, in the context of the exploration of electrophysiological signals for emotional states detection, showed the existence of differentiated states in the data that represent the ECG signal of a subject. To deal with this intra-class differences the proposed method includes in the “model” (as mentioned in Section C.3) single heartbeats randomly selected from the whole ECG signal sample.

The reported results are averages over 50 runs. In each run, we partition the set of heartbeats of each subject into two mutually exclusive subsets: one of these subsets is used to form the training data set $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2,..., \mathcal{X}_K\}$, while the other is used to build the test waveforms. We consider several values for n (the length of the “model” strings) as well as for m (the length of the test waveforms).

C.4.2 Identification Results

The results for the identification experiment, which are depicted in Fig. C.4, show that the proposed method achieves 100% accuracy for $m = 12$ and $n = 13$ or $n = 20$. This is better than the results reported in [20] over the same dataset. The approaches in [1], [19], [5], were not tested on the same dataset, so the results are not directly comparable. Notice that using only $m = 5$ waveforms for the test patterns, we already reach an accuracy around 99.5%. As expected, the accuracy increases both with n and m.

C.4.3 Authentication Results

Regarding verification (authentication) three different test were made. The first test follows the model shown in Fig. C.3. The results, which are depicted in Fig. C.5 (a), show that the proposed method achieves an overall equal error rate EER $\approx 6\%$. Notice that one can lower the error rate using lower threshold values but then the system will reject more legitmate users. However, it is possible to use lower threshold values if we use a different value for each subject (user-tuned thresholds).

The second test also follows the model shown in Fig. C.3 but now the threshold is user-tuned. An equal error rate (EER) was computed for each subject and then an average EER is reported. The test results presented in Table C.1 show that the proposed method outperforms fiducial approaches results reported in [9] and [16], over the same dataset.

Reference	**Feature**	**EER**
Oliveira and Fred [16]	Fiducial (1-NN classifier)	8.0 %
Gamboa [9]	Fiducial (user tuned)	1.7 %
Proposed method	Non-fiducial (user tuned)	1.1 %

Table C.1: Comparison of verification related work results with our method, over the same dataset.

On the last verification test, we evaluate the combination of multiple source acquisition signals, classified by a bank of classifiers with the same structure of the first test, shown in Fig. C.3, and a final decision made according to the majority voting criterion. Given a test sample (of length $m = 12$), it was decomposed in 64 different ways into samples of length $m = 6$ which were classified by a bank of 64 classifiers using the same threshold level and the same database. The results in Fig. C.5 (b) show that this multiple classifier strategy doesn't improve the performance.

C.5 Summary and Conclusions

We have presented a method for personal identification and authentication from one-lead ECG signals which involves no explicit feature extraction other than 8 bit uniform quantization of the waveforms. The classifier is based on the Ziv-Merhav cross parsing (ZMCP) algorithm, which is an estimator of the algorithmic cross-complexity [3], used to measure the similarity between the model waveforms and the test waveforms. Experiments carried out on a dataset with 19 healthy subjects, for whom the existence of differentiated states in the ECG data of a subject has been shown [14], showed that our method achieves 100% accuracy in recognition (identification) and an average equal error rate close to 1.1% in verification (authentication) tasks. Although further experiments, on other datasets,

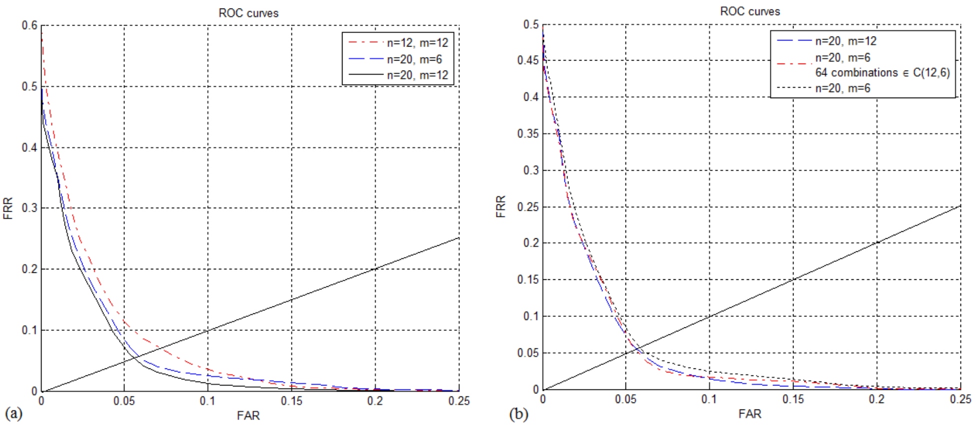


Figure C.5: ROC curves for the verification task (the solid straight line has slope 1, for reference purposes). Left plot: results for different n and m values; notice the improvement with the increase of m and n. The best equal error rate (EER) is close to 6%. Right plot: results for single classifiers versus a bank of 64 classifiers with the same structure for combination of multiple source acquisition signals.

are needed to assess the relative performance of the proposed method, with respect to other state-of-the-art techniques, these results demonstrated the validity of our approach as a tool for personal identification and authentication, and of the ECG signal as a viable biometric. Future work will include tests with the Max-Lloyd quantizer and further evaluation of our method when used in an adaptive way for authentication purposes with continuous biometrics systems [8].

Acknowledgments

We acknowledge the following financial support: the FET programme, within the EU FP7, under the SIMBAD project (contract 213250); Fundação para a Ciência e Tecnologia (FCT), under grants PTDC/EEA-TEL/72572/2006 and QREN 3475; Departamento de Engenharia de Electrónica e Telecomunicações e de Computadores, Instituto Superior de Engenharia de Lisboa.

References

[1] L. Biel, O. Pettersson, L. Philipson, and P. Wide. ECG analysis: a new approach in human identification. *IEEE Transactions on Instrumentation and Measurement*, 50(3):808–812, June 2001.

[2] Nikolaos V. Boulgouris, Konstantinos N. Plataniotis, and Evangelia Micheli-Tzanakou. *Biometrics: theory, methods, and applications*. Wiley-IEEE Press, 2009.

[3] Daniele Cerra and Mihai Datcu. Algorithmic Cross-Complexity and Relative Complexity. *2009 Data Compression Conference*, pages 342–351, March 2009.

[4] ADC Chan and MM Hamdy. Wavelet distance measure for person identification using electrocardiograms. *Instrumentation and Measurement, IEEE Transactions on*, 57(2):248–253, 2008.

[5] Chuang-Chien Chiu, Chou-Min Chuang, and Chih-Yu Hsu. A Novel Personal Identity Verification Approach Using a Discrete Wavelet Transform of the ECG Signal. In *2008 International Conference on Multimedia and Ubiquitous Engineering (MUE 2008)*, pages 201–206. IEEE, 2008.

[6] David Pereira Coutinho and Mário A. T. Figueiredo. Information Theoretic Text Classification Using the Ziv-Merhav Method. In *Pattern Recognition and Image Analysis. Springer Berlin Heidelberg*, volume 1, pages 355–362, 2005.

[7] J. Cunha, Bernardo Cunha, William Xavier, Nuno Ferreira, and A. Pereira. Vital-Jacket: A wearable wireless vital signs monitor for patients mobility. In *Proceedings of the Avantex Symposium*, 2007.

[8] I. G. Damousis, D. Tzovaras, and E. Bekiaris. Unobtrusive multimodal biometric authentication: The HUMABIO project concept. *EURASIP Journal on Advances in Signal Processing*, 2008:110, 2008.

[9] Hugo Gamboa. *Multi-Modal Behavioral Biometrics Based on HCI and Electrophysiology*. Phd thesis, Universidade Técnica de Lisboa, Instituto Superior Técnico, 2008.

[10] Steven A. Israel, John M. Irvine, Andrew Cheng, Mark D. Wiederhold, and Brenda K. Wiederhold. ECG to identify individuals. *Pattern Recognition*, 38(1):133–142, January 2005.

[11] A. K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, January 2004.

[12] Pablo Laguna, Roger G. Mark, A. Goldberg, and George B. Moody. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. In *Computers in Cardiology 1997*, pages 673–676. IEEE, 1997.

[13] Vladimir Leonov, Tom Torfs, Inge Doms, Refet Firat Yazicioglu, Ziyang Wang, Chris Van Hoof, and Ruud Vullers. Wireless body-powered electrocardiography shirt. In *Proceedings of the 3rd European Conference on Smart Systems Integration*, pages 307–314, 2009.

[14] Liliana Medina and Ana Fred. Genetic Algorithm for Clustering Temporal Data-Application to the Detection of Stress from ECG Signals. In *Proceedings of 2nd International Conference on Agents and Artificial Intelligence (ICAART)*, pages 135–142, 2010.

[15] Mark Nelson and Jean-loup Gailly. *The Data Compression Book*. M&T Books, New York, 2nd editio edition, 1995.

[16] Carla Oliveira and Ana L. N. Fred. ECG-based Authentication-Bayesian vs. Nearest Neighbour Classifiers. In *Proceedings of International Conference on Bio-inspired Systems and Signal Processing - Biosignals - INSTICC*, pages 163–168, 2009.

[17] Arun Abraham Ross, Karthik Nandakumar, and Anil K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, 2006.

[18] D. Salomon. *Data Compression. The complete reference*. Springer-Verlag New York, 3rd edition, 2004.

[19] T. W. Shen, W. J. Tompkins, and Y. H. Hu. One-lead ECG for identity verification. In *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, pages 62–63, 2002.

[20] Hugo Silva, Hugo Gamboa, and Ana Fred. One Lead ECG Based Personal Identification with Feature Subspace Ensembles. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2007)*, pages 770–783. Springer-Verlag Berlin, Heidelberg, 2007.

[21] H. H. So and K. L. Chan. Development of QRS detection method for real-time ambulatory cardiac monitor. In *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, volume 289, pages 289–292, 1997.

[22] Yongjin Wang, Foteini Agrafioti, Dimitrios Hatzinakos, and Konstantinos N. Plataniotis. Analysis of Human Electrocardiogram for Biometric Recognition. *EURASIP Journal on Advances in Signal Processing*, 2008(1):19, 2008.

[23] John Wright, Yi Ma, Yangyu Tao, Zhouchen Lin, and Heung-Yeung Shum. Classification via Minimum Incremental Coding Length. *SIAM Journal on Imaging Sciences*, 2(2):367–395, January 2009.

[24] J. Ziv and a. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977.

[25] J. Ziv and a. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, September 1978.

[26] J Ziv and N Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39(4):1270–1279, 1993.

Appendix D

Novel Fiducial and Non-fiducial Approaches to ECG-based Biometric Systems

Paper D

Novel Fiducial and Non-fiducial Approaches to ECG-based Biometric Systems

David Pereira Coutinho, Hugo Silva, Hugo Gamboa, Ana Fred and
Mário Figueiredo

In *Journal of IET Biometrics*, 2(2), pages 64–75, 2013

Abstract

The electrocardiogram (ECG) is a non-invasive and widely used technique for cardiac electrophysiological assessment. Although the ECG has traditionally only been used for functional diagnostic and evaluation, several advances in electrophysiological sensing have made available robust signal acquisition devices, particularly suited for ambulatory conditions, widening its range of applications. In particular, recent work has shown the potential of the ECG as a biometric trait, both for human identification and authentication. This study sets the ground for an ECG-based real-time biometric system. We describe an experimental setup and the evaluation of new fiducial and non-fiducial approaches, including data acquisition, signal processing, feature extraction and analysis, and classification methodologies, showing the applicability of the ECG as a real-time biometric. Performance evaluation was done in clinical-grade ECG recording from 51 healthy control individuals (of a publicly available benchmark dataset) as well as on data collected from 26 healthy volunteers performing computer activities without any posture or motion limitations, thus simulating a regular computer usage scenario.

D.1 Originality and Contributions

In this paper, we present an original work addressing the use of electrocardiographic (ECG) signals for human recognition purposes, through fiducial and non-fiducial approaches. So far, the ECG has been used mostly for cardiac electrophysiological assessment in clinical applications. Moreover, several advances in electrophysiological sensing have made available robust signal acquisition devices, particularly suited for ambulatory conditions, widening its range of applications.

State of the art work has shown the potential of the ECG as a behavioral biometric trait, both for identification and authentication. The ECG is particularly interesting to complement other biometric features in a multibiometric system. Unlike other traits, the ECG provides intrinsic aliveness verification, as well as the detection of different emotional states, among other advantages.

We expand the state of the art, by proposing: a fiducial method that recurs to a limited number of patterns and features; a non-fiducial approach based on information theoretical methods; and by devising the two approaches targeting real-time operation. Furthermore, unlike prior-art work, which uses ECG collected in a rested position clinical setting, we also performed data acquisition and system evaluation with data collected in a real-world setting (HiMotion dataset). Specifically, ECG was recorded from 26 healthy volunteers performing activities at the computer, without any posture or motion limitations to simulate a regular computer usage scenario.

Experiments were done on the publicly available PTB dataset and on the HiMotion dataset. Test results further enhance the applicability of the ECG signal as a biometric trait, and confirm its biometric potential even on data acquired in unrestrained scenarios.

In summary, our contributions are as follows: we present a novel non-fiducial ECG biometrics method based on the Ziv-Merhav cross parsing length; we present a fiducial method based on template matching that has a straightforward implementation in real-world deployment scenarios, and that is used for baseline performance comparison and demonstration of the capabilities of the non-fiducial method; we provide a sensitivity analysis on the influence of the number of templates considered for the training (models) and testing sets on the overall performance of the classifiers; besides the evaluation of our algorithms in data collected with a clinical/rest scenario (in our case the PTB database), which is the typical approach found in the reference literature, we also use data collected in a context that has a direct parallel with a possible real-world application scenario; finally, we provide a brief literature review, that highlights the permanence of the ECG template in terms of its biometric recognition performance.

D.2 Introduction

The *electrocardiogram* (ECG) is a technique to acquire, store and/or visualize the signals produced by the electrical activity of the heart, commonly used as a clinical diagnostic tool for the evaluation of the cardiac function. Each cardiac heartbeat cycle depicts the evolution of this electrical activity over time, and is characterized by a collection of *complexes P-QRS-T* [10]. Waveform features related to latency and amplitude can be extracted from each complex or from the relation among different complexes[1] and provide useful information for ECG signal analysis in clinical applications.

Recently, the ECG signal has emerged as a biometric trait, exploting a physiological feature that exists on every human. There is strong evidence that the ECG is sufficiently discriminative to identify individuals from a relatively large population. Moreover, the ECG allows intrinsic liveliness verification, personal identification [5] and authentication [9], as well as the detection of different stress or emotional states [27]. The ECG is a behavioral biometric trait that can be used together with other types of biometric features [32] in a multimodal biometric system [6].

In previous research work [41], the ECG-based biometric methods were divided in two classes: fiducial and non-fiducial. Fiducial methods use points of interest within the heartbeat wave (*P-QRS-T complex*), which are then used to extract latency and amplitude features. During the enrollment/training phase, these features are extracted from the ECG signals of the individuals and stored in a database. During the identification or authentication phase, the features extracted from the user ECG waveforms are compared with those stored in the database and a decision is made.

On the other hand, non-fiducial techniques aim at extracting discriminative information from the ECG waveform without having to localize fiducial points, making feature extraction in the frequency domain, or even using the values of a global pattern from several heartbeat waves as features. Moreover, there are some methods which combine these two different approaches and are called partially fiducial. In this paper, we address the use of the ECG signal for human identification and authentication purposes through both fiducial and non-fiducial approaches.

The proposed approaches have been benchmarked using ECG recordings from 51 healthy control individuals of the PTB dataset [18], and also on ECG recordings collected from 26 heathy volunteers, while performing activities at the computer, without any posture or motion limitations to simulate a regular computer usage scenario. The main contributions of our work are the demonstration of a novel string matching non-fiducial method as a viable alternative to fiducial methods (here used as a baseline for comparison), and the further extension of the state of the art with respect to earlier references, by adopting an unconstrained and working environment instead of a rest position, and by using data from a single sensor lead.

The structure of the paper is as follows: Section D.3 provides a brief introduction to the problem; Section D.4 explains the data acquisition and signal processing procedures; Section D.5 describes the proposed fiducial and non-fiducial approaches, detailing the feature extraction step, the classification method, the quantization and string matching steps; Section D.6 presents the analysis methodology, classification method and experimental results; finally Section D.7 summarizes the main conclusions.

[1] examples of features that relate different complexes are the duration of the *PR* and *ST* segments, and the *QT* interval.

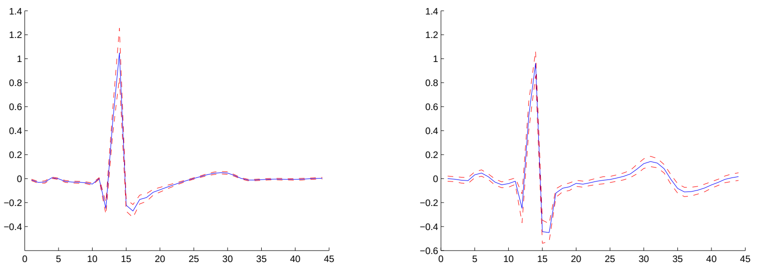
D.3 Using the ECG for Human Biometrics

D.3.1 The ECG as a Biometric

As measured on the chest surface, the human ECG signal is directly related to the physiology of each individual. Such information varies amongst individuals due to factors such as skin conductance, body mass, congenital disorders, genetic singularities, position, shape, and size of the heart and chest cavity, among others. Regardless of what factors originate these differences, the fact that there are subject-specific physiologic features in the ECG signal suggests its applicability to the context of biometric systems.

The ECG is highly correlated with the subject's physical state and condition, as well as with his/her emotional state [27]; this fact makes the ECG a very appealing biometric modality in terms of integrity, since the system can be designed to require physiological activity in a condition similar to the one required during the enrollment phase. Furthermore, the ECG is not easily spoofed or masqueraded, unlike other methods [17], [26].

Figure D.1 illustrates the ECG signals acquired on two different subjects (using the same experimental setup and procedures). The signals are normalized and, as it is easily observed, both waveforms evidence clear commonalities but also some distinct features.



(a) Subject A

(b) Subject B

Figure D.1: Normalized ECG readings from two different subjects acquired as described in Section D.4.1. The solid and dashed lines correspond respectively to the mean wave and standard deviation computed over the full set of patterns for each of the subjects.

As a way of further improving system security, considerable work has been devoted to the application of biometric methods in a continuous approach [23], [3], [35]. The motivation for continuous biometrics is twofold: on one hand it addresses the issue of guaranteeing that the user remains the same throughout a session, after having been initially granted access; on the other hand it can be used as an additional source for security enhancement and/or redundancy in a multimodal approach [28], [6], [21], [32].

In a continuous biometric framework, the ECG plays an important role; not only is it bound to electrophysiological features of the individual and has been shown to perform accurately for human identification[20], [34], but it can also be easily collected continuously. Furthermore, a direct application of the ECG as a biometrics is specially simple in systems where this signal is already measured, as is the case of medical applications, as a way of providing automated patient recognition ability.

In the past, ECG acquisition systems were cumbersome and somewhat invasive. Nowadays, ECG signals are easily collected; recent advances in body area networking and electrophysiological signal acquisition devices have brought great improvements in terms of autonomy and size, facilitating the acquisition process, even in ambulatory scenarios [29], [13], [36], [25]. Moreover, advances in system integration are yielding autonomous ECG systems that can be unobtrusively worn by the subjects [24], [13], [37].

D.3.2 State of the Art

Biel et al. [5] pioneered the use of the ECG as a biometric for personal identification. They initially used a 12-lead ECG, which requires meticulous and unpractical placement of the electrodes on each person, but ended up concluding that one lead was enough. Using a proprietary equipment from SIEMENS, 30 fiducial features were extracted; a feature selection algorithm allowed concluding that the best results were obtained with 10 of these features, and using a principal component analysis (PCA) of each class. The purpose was to identify 20 subjects at rest, a task on which they achieved 100% accuracy.

Using a predetermined group of 20 subjects, selected from the MIT/BIH ECG database [18], Shen et al. performed experiments targeting ECG-based identity verification [34]. Through a template matching technique, a 95% accuracy was obtained, while using a neural network classifier lead to 80% accuracy; a method combining both techniques achieved 100% identity verification accuracy.

Using a setup for palm-based ECG measurement, experiments are reported in [33] on data collected from 168 subjects in a resting scenario. By combining a template matching method with a neural network, for predetermined groups of 10, 20, and 50 subjects, 100% recognition rates were reported; for a predetermined group of 100 subjects, 96% recognition rate was achieved; finally, a recognition rate of 95.3% was reported for the complete set of 168 subjects [33].

In [20], experiments were performed on data collected from 29 subjects while performing a set of 7 activities. The recordings were performed on the chest and neck, and 12 temporal features extracted from the signal were used. Using standard linear discriminant analysis (LDA), individual waveforms are classified and mapped to the identity of the subject by a majority voting scheme. The authors report an accuracy of 100% in subject identification.

A method for feature extraction from the one-lead ECG signal, based on a combination of autocorrelation analysis (AC) with the discrete cosine transform (DCT), was introduced in [41]. This method does not require segmentation of the ECG signal into heartbeats, with only the *R* peak detection being needed for the *QRS* window identification. In a subject identification task (on a subset of 13 subjects from the MIT-BIH dataset), the authors used a nearest-neighbor classifier based based on the normalized Euclidean distance between feature vectors, and reported a recognition rate of 97.8%.

A system based on a 3 step feature extraction method was introduced in [9]. The method uses the *QRS-complex* detection algorithm proposed in [39], and the discrete wavelet transform to extract signal features. A nearest-neighbor classifier based on the Euclidean distance between feature vectors is used. On 35 subjects from the QT database [18], the authors report an accuracy of 100% for the identification task and verification rates of 0.83% of false acceptance rate (FAR) and 0.86% of false rejection rate (FRR).

Human identification based on an ECG acquired from the fingers is possible, as shown in [8]. The authors introduced a simple non-clinical data acquisition setup based on 2 button electrodes, which the subjects hold between the pads of their thumb and index fingers. The *P-QRS-T complexes* are detected and temporally aligned, in order to compute an average ECG; the proposed classifier uses a distance measure based on wavelet coefficients. A 89% identification accuracy was achieved on a dataset of 50 individuals.

Table D.1 summarizes the main characteristics and results of the several approaches reviewed in the previous paragraphs; more details on each method may of course be found on the corresponding publications.

Table D.1: Comparison of related work with our methods. The accuracy (Accur.) values shown are the reported results for person identification, on databases of different sizes (Subjs.)

Ref.	**Feature**	**Method**	**Subjs.**	**Accur.**
[5]	Fiducial	PCA	20	100%
[34]	Fiducial	Template matching + DBNN	20	100%
[20]	Fiducial	LDA	29	100 %
[41]	Non-fiducial	AC/DCT+KNN	13	97.8%
[9]	Non-fiducial	Wavelet Distance	35	100%
[8]	Non-fiducial	(fingers) Wavelet Distance	50	89%
Proposed	Fiducial	Mean Waves + 1NN	51	99.85%
Proposed	Non-fiducial	String Matching + 1NN	51	99.39%

D.3.3 Fiducial Versus Non-fiducial Approaches

In a broad sense, one can say there are two different approaches in the literature concerning feature extraction from the ECG: fiducial [5], [34], [20], and non-fiducial [8], [9]. Fiducial methods use points of interest within a single heartbeat waveform, such as local maxima or minima; these points are used as reference to allow the definition of several time and amplitude features (see Fig. D.5). Non-fiducial techniques extract discriminative information from the ECG waveform without localizing fiducial points. In this case, a global pattern from several heartbeat waveforms may be used as a feature. Some methods which combine these two approaches are called partially fiducial [41] (e.g., they use only the *R* peak as a reference for segmentation of the heartbeat waveforms).

Our work focuses on the potential of human identification and authentication using a reduced number of heartbeat waveforms, with the purpose of continuous biometrics applications. For both the fiducial [38] and non-fiducial [12] approaches, we evaluate the recognition rate of an average heartbeat waveform in terms of identification and authentication discriminative potential. Both approaches use the same ECG preprocessing and acquisition procedures.

D.4 Data Acquisition and Signal Processing

D.4.1 Data Acquisition

In this paper, two databases were used to assess the performance of the proposed methods: (a) the HiMotion database, in which data was collected by the authors from healthy volunteers (students); and (b) the *PTB Diagnostic ECG Database*, prepared for the PhysioNet[2]; details about the acquisition conditions and the preprocessing can be found at the PhysioNet site[2].

Regarding the HiMotion database, data was collected from 26 volunteers (18 males and 8 females, between the ages of 18 and 31), who willingly participated in individual sessions (one per subject), during the course of which their ECG signal was recorded. Unlike the conventional settings found in the literature, and in a similar way like the one adopted by Riera et al. in [30] where the acquisition is performed in a resting position, in each session the subject was asked to complete, at a computer, a task requiring mental concentration and designed for an average completion time of 10 minutes. We refer the reader to [15] for further details on the database and experimental setting.

The task was designed in such way that the subject only used the mouse to interact with the computer. To motivate the subjects commitment to the test, a performance score was computed and assigned to each subject. Unlike the standard 12-lead ECG recording setup, which involves six sensors placed on the chest area and six other placed on the limbs, we used a one-lead surface mount setup, which has been shown to suffice for ECG-based biometrics [5]. Acquisition was performed at a $256Hz$ sampling rate using a ProComp2 encoder, and a gain 50 local differential triode (Figure D.2 on the left), with $2cm$ spacing between electrodes, and channel bandwidth of $0.05 - 1kHz$, both from Thought Technology Ltd.

For sensor placement, the V_2 precordial derivation was chosen, located on the fourth intercostal space in the mid clavicular line, at the right of the sternum (Figure D.2 on the right). Ten-20 conductive paste was used to improve conductivity; for the same purpose, prior to the sensor placement, the selected area was prepared with abrasive gel.

Prior to initiating the task, subjects were equipped with the sensor and placed in front of the computer in a sitting position. No limitations on posture or motion during the activity were imposed. Figure D.3 shows the apparatus involved in the acquisition process. Other sensors and systems are visible in Figures D.3(a) and D.3(b) since the ECG acquisition was performed in the wider context of a project on multi-modal behavior biometrics and user authentication.

[2] http://www.physionet.org/physiobank/database/ptbdb/

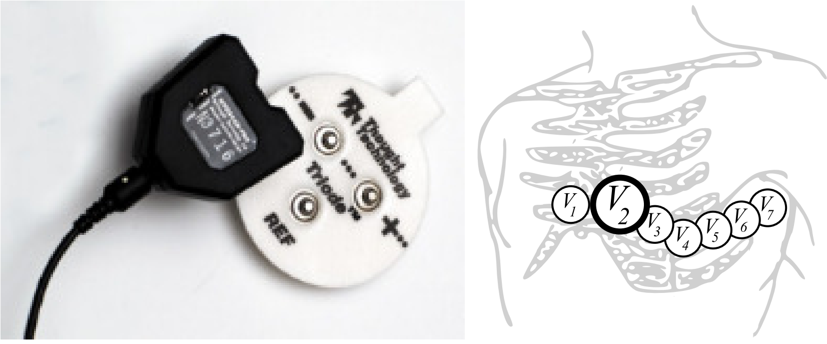
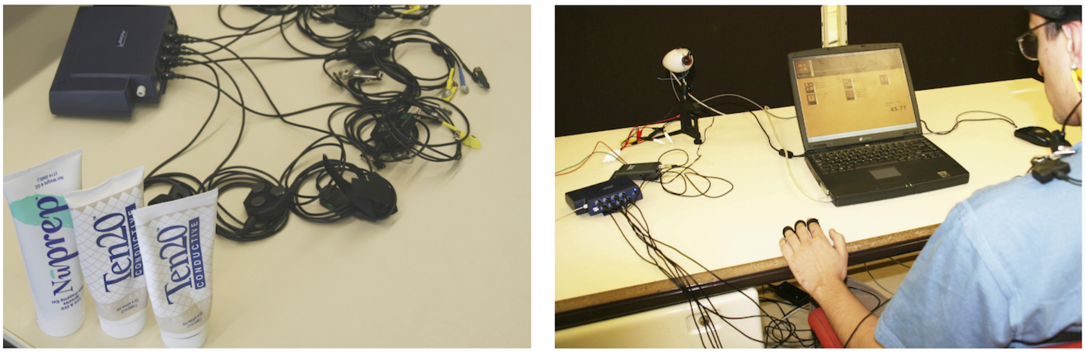


Figure D.2: ECG acquisition sensor and lead V_2 sensor placement used in the experimental setup.



(a) Acquisition device and accessories.

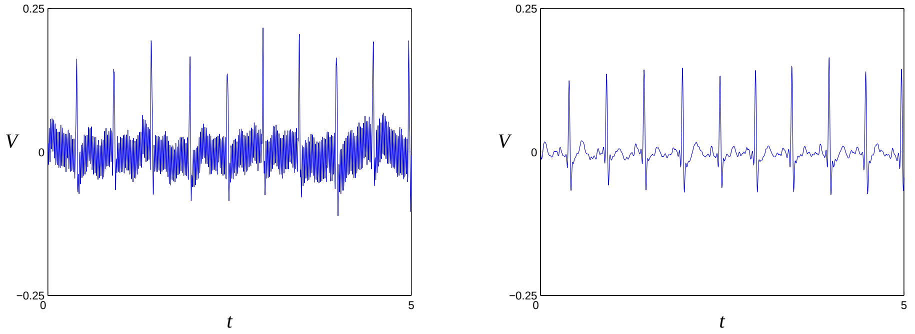
(b) Subject placement and apparatus.

Figure D.3: ECG acquisition device, accessories, and apparatus.

D.4.2 Signal Processing

Figure D.4(a) shows 5 seconds of raw ECG signal, acquired under the conditions described in Section D.4.1. As we can see, the interest areas of the signal are immersed in noise, with only the *R* peaks being clearly noticeable. Mainly two types of noise are present in the signal: low frequency noise, which we can observe in Figure D.4(a) expressed as the baseline fluctuation of the ECG signal trace; wide-band noise (usually modeled as white), with its characteristic appearance. These perturbations are introduced by electrical/magnetic fields and by the acquisition hardware. In order to improve the ECG waveform, we filter the noise components from the signal. To identify a general region of interest, we took as a reference 60 seconds of ECG readings at the V_2 lead from control patient 104 of the PTB ECG diagnostics database [18].

To filter the acquired signals, we used a 4^{th} order Butterworth filter, since it presents a stable response at the frequency band limits. Our first approach was a bandpass on the $1-30Hz$ frequency range; however, in some cases, this choice still left some low frequency baseline fluctuation. Since we needed a filter that could be adapted to all users, we narrowed the band to the $2-30Hz$ frequency range, which proved to be more robust. Further improvements were obtained using a zero-phase



(a) Raw signal.

(b) Filtered signal.

Figure D.4: Raw and filtered ECG signals example (approximately 5 seconds of acquired signal from one of the subjects).

forward-reverse scheme [19], in which the signal is filtered in two steps; first, the signal is directly passed through the filter; the resulting signal is time-reversed and passed through the filter once more. Figure D.4(b) illustrates the signal of Figure D.4(a) filtered using the adopted approach.

After filtering, each individual heartbeat waveform was segmented from the full recording using a derivation of the *multiplication of the backward distance* (MOBD) algorithm [22], [40], [42]. All segmented heartbeat waveforms were aligned, by their *R* peaks, in segments of equal duration. From the resulting collection of ECG heartbeat waveforms, the non-overlapping mean wave for groups of 10 consecutive heartbeat waveforms was computed, to minimize the effect of outliers. Finally, a labeled database was built, in which each pattern corresponds to a mean wave; this procedure was followed for both the HiMotion and PTB databases.

D.5 Proposed Approaches

D.5.1 Fiducial

As described in Section D.3.3, fiducial approaches rely on notable points in the ECG patterns. Our fiducial approach is based on features extracted from the global patterns of several heartbeat waves, to convert the raw discrete-time ECG signal into a set of latency and amplitude features. As described in Section D.2, the ECG heartbeat waveform is characterized by a collection of complexes identified as *P*-*QRS*-*T*. No time limit was imposed to complete the task, and therefore the heartbeat waveform collection of each subject in the database was truncated at approximately 6 minutes, which corresponds to the fastest completion time over all the subjects.

A wide range of features can be used to characterize the ECG signal for human identification [22], [10], [14]. In our approach, for each mean waveform, the *R* peak is taken as the reference complex (with t_r being the reference latency), and from this the remaining complexes are determined. The Q and *S* complexes are determined as the point with minimum amplitude value found respectively at the left and right of the R peak; the P and T complexes are determined as the points with maximum amplitude value found respectively at the left and right of the R peak. From this information, 8 latency and amplitude features were extracted from the complexes, along with a sub-sampling of the waveform itself. This resulted in a feature space of dimension $d = 53$: 4 latency features (P_t, Q_t, S_t, T_t), 4 amplitude features (P_a, Q_a, S_a, T_a), and 45 amplitude values resulting from sub-sampling the mean waveform. Figure D.5 depicts the features extracted from each complex; the remaining features were determined by reducing the signal sampling rate to $64Hz$.

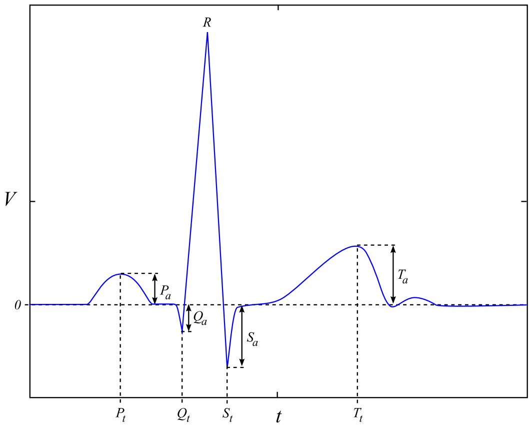


Figure D.5: Features measured from the ECG waveform. After the filtering process the ECG baseline is fairly zero centered, and therefore the amplitude measurements are taken with respect to the zero centered line. The waveform segments for each subject all have the same length, and therefore the temporal measurements are relative measures taken with respect to the R peak.

D.5.2 Non-fiducial

As mentioned in Section D.3.3, non-fiducial methods base their decision directly on the waveform, without extracting intermediate features. In this paper, we propose a novel and simple approach based on information theoretic tools, which uses quantization to convert the ECG discrete-time analog signal values into a sequence of symbols (a string), followed by string matching as a tool for classification.

D.5.2.1 Quantization

The simplest approach to convert a set of single heartbeat waveforms into a set of strings is to apply N-bit uniform quantization, which produces sequences of symbols (strings) from an alphabet with 2^N symbols. Thus, a collection of heartbeat waveforms is transformed into a collection of strings. Therefore all the tools developed for text classification and string matching can be applied to ECG classification.

Quantization with less then 8 bits was considered in early experiments, but discarded because the resulting performance was lower than with 8-bit quantization. Higher values were not considered essentially because the performance obtained with 8 bits was considered satisfactory. Despite the information loss due to the quantization process, our experimental results show that enough discriminative information is preserved.

To improve the discriminative power of the system, we propose to use non-uniform quantizers adapted to each user. In particular, we adopt optimal quantizers obtained by the well-known Lloyd-Max algorithm, which minimizes the MSQE (mean squared quantization error). Lloyd-Max quantization has been previously used in ECG compression for transmission purposes [31]. Our proposal is that a Lloyd-Max quantizer is obtained for each user after the enrollment process, and the selected heartbeat waveforms (user's model) be encoded with that user-tuned quantizer. Both the user-tuned quantizer and the encoded waveforms (model) are stored in the database, so that it can be used later. During the verification process, the selected heartbeat waveforms are encoded with the Lloyd-Max quantizer associated to the subject that the user claims to be, and then compared with that subject's model.

D.5.2.2 String Matching

The novel non-fiducial method proposed in this paper is grounded in information theoretic text classification concepts and tools, namely the concept of cross complexity introduced in [7] and the Lempel-Ziv based cross parsing algorithm described in [45]. The Lempel-Ziv (LZ) algorithm [44] is a well-known tool for text compression, which in recent years has also been used for text classification purposes [4] [7]. In particular, it was shown that the Ziv-Merhav (ZM) method, originally proposed for measuring relative entropy [45], achieves state of the art performance in a specific text classification task [11].

Ziv and Merhav [45] proposed an empirical divergence estimator between two sequences $\mathbf{z}$ and $\mathbf{x}$, based on two LZ-type parsing algorithms: the incremental LZ parsing (LZ78) which is a self parsing procedure of a sequence $\mathbf{z}$ and the cross parsing (LZ78 parsing variation) which is a sequential parsing of a sequence $\mathbf{z}$ with respect to another sequence $\mathbf{x}$. Roughly speaking, we can see the self parsing length of a given sequence $\mathbf{z}$, denoted as $c(\mathbf{z})$, as a measure of the complexity of the sequence $\mathbf{z}$. Similarly, the sequential parsing length of a sequence $\mathbf{z}$, with respect to a given sequence $\mathbf{x}$, denoted as $c(\mathbf{z}|\mathbf{x})$, can be seen as the code-length obtained when coding $\mathbf{z}$ using a model for $\mathbf{x}$ (cross parsing), thus providing an estimate of the cross complexity [7]. It is expectable that the cross complexity is low when the two sequences are very similar, which is the key idea behind the use of cross parsing for classification [11]. In this paper, we apply this idea to the problem of ECG biometrics.

An implementation of ZM cross parsing (ZMCP), as a component of the ZM method for relative entropy estimation, was proposed in [11], based on a modified LZ77 [44] algorithm in which the dictionary is static and only the lookahead buffer slides over the input sequence (see Figure D.6). This very same implementation of cross parsing, using a 64 Kbyte dictionary and a 256 byte look ahead buffer, was used in the non-fiducial experiments to compute the proposed string similarity measure $c(\mathbf{z}|\mathbf{x})$. However to allow for the definition of a threshold level between 0 and 1 for authentication purposes, we propose the use of a normalized version,

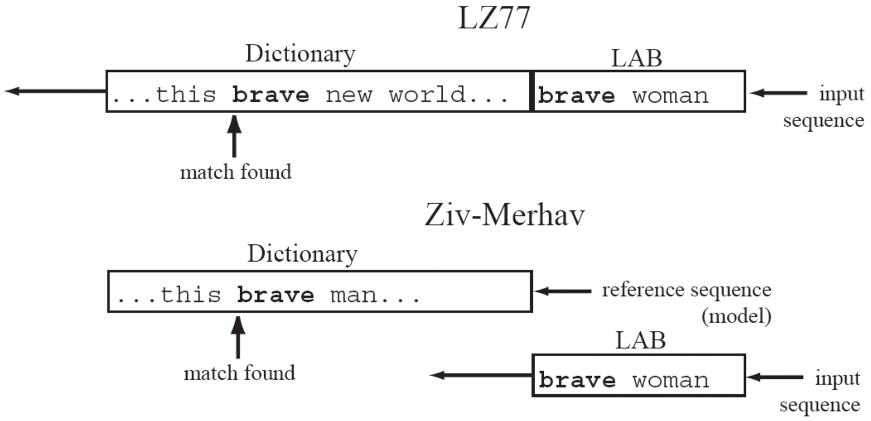


Figure D.6: The original LZ77 algorithm uses a sliding window over the input sequence to encode and update the dictionary; in the used cross parsing algorithm implementation, the dictionary is static and only the lookahead buffer (LAB) slides over the input sequence.

$$C(\mathbf{z}|\mathbf{x}) = \frac{c(\mathbf{z}|\mathbf{x})}{|\,\mathbf{z}\,|},$$

which yields values in this range. In this definition, $|\,\mathbf{z}\,|$ is the (byte) length of the sequence $\mathbf{z}$. Notice that when the strings are very different, the estimated cross complexity will be close to $|\,\mathbf{z}\,|$, making $C(\mathbf{z}|\mathbf{x}) = 1$. For very similar strings, the estimated cross complexity will be low and thus C will be close to zero.

Therefore, given an ECG sample string $\mathbf{z}$, built from concatenated single heartbeat strings of an unknown subject, its identity is estimated by computing $C(\mathbf{z}|\mathbf{x})$ for every possible known subject $\mathbf{x}$ and deciding for the one achieving the lowest value. In other words, the sample will be classified as belonging to the subject that leads to its shortest description.

D.5.3 Classification

For identification, both the fiducial and non-fiducial approaches use a classification method based on a 1-NN (nearest neighbor) criterion, where the comparison of the enrollment ECG templates from the system database against the identification ECG is based on the computed value of a distance measure.

Consider a collection of training patterns partitioned into K classes (the set of subjects to be identified): $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_K\}$. For each subject/class k, $\mathcal{X}_k = (\mathcal{P}_{k,1}, \mathcal{P}_{k,2}, \dots, \mathcal{P}_{k,n})$ denotes a set of n patterns, each corresponding to an averaged single heartbeat waveform (as previously mentioned in Section D.4.2). Each string $\mathbf{x}_k = (\mathbf{p}_{k,1}, \mathbf{p}_{k,2}, \dots, \mathbf{p}_{k,n})$ is formed by concatenating the n quantized patterns of subject k; string $\mathbf{x}_k$ is, in some sense, a “model” representing the shape of the heartbeats of subject k.

The 1-NN based identification of a test sample (string) $\mathbf{z}$, built from m heartbeats obtained from the test set of an unknown subject, in one of a set of K classes, given the subject models $\mathbf{x}_k$ per class k, is given by

$$\hat{k}(\mathbf{z}) = \arg\min_{k\in\{1,\ldots,K\}} C(\mathbf{z}|\mathbf{x}_k).$$

In the fiducial approach, we apply the Euclidean distance metric to compute the distance between the features extracted from the test sample and the features of each of the models in the database, while in the non-fiducial approach, we apply the proposed string similarity measure (ZMCP) to evaluate the similarity between the test sample string and each user's model string from the database. An overview of the non-fiducial ECG-based human identification system is depicted in Figure D.7, where we can see that the comparison of the enrollment ECG templates from the system database against the identification ECG is based on the cross parsing computed value.

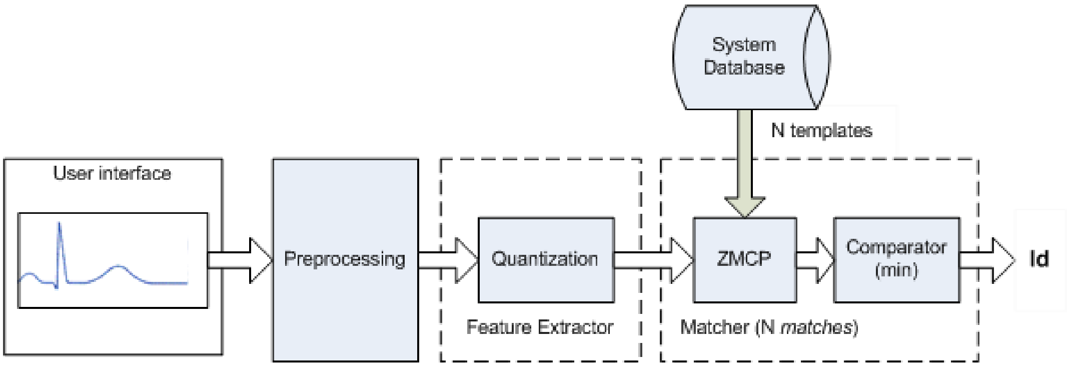


Figure D.7: Non-fiducial ECG-based human identification system overview. The classification method is based on a 1-NN (nearest neighbor) classifier, which uses the Ziv-Merhav cross parsing (ZMCP) length as a string similarity measure for the decision process.

For authentication, the classifier also uses the Euclidean distance in the fiducial case, and the defined string similarity measure (ZMCP) in the non-fiducial approach, to determine the distance between a given unknown pattern $\mathbf{z}$ and the known template $\mathbf{x}_k$ in the database for whom the user claims to be, and compares it to a threshold. In our approach, we adopt a user-tuned threshold, previously established during the enrollment process for that particular user.

Regarding the training phase, whenever a new user is added to the system, neither the fiducial nor the non-fiducial approaches require access to the dataset with all the previously enrolled users. For a given new user enrolled in the system, a new training pattern $\mathcal{X}_{K+1}$ will simply be added to the collection (system database), that is $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2,..., \mathcal{X}_K\} \cup \{\mathcal{X}_{K+1}\}$, and the number of classes updated $K = K + 1$. A new feature vector (template) is generated from the information in the acquired ECG signal (see Figure D.7); afterwards, we perform the raw data preprocessing, mean waves computation, and feature extraction or quantization, plus string concatenation (the former for the fiducial approach and the later for the non-fiducial approach, as explained in the previous sections). Applying the classification rule during the test phase implies only the comparison of the acquired test pattern with the stored patterns in the database, using the adopted distance or similarity measure.

D.6 Experimental Results

D.6.1 Databases

As mentioned in Section D.4, our work uses two ECG databases, the HiMotion database, which has data from 26 subjects collected in an unrestrained scenario while performing a regular computer-based task, and the PTB diagnostics dataset, which has data from 51 healthy subjects collected in a clinical scenario.

In order to have the same amount of information for all subjects, a selection of a fixed number of patterns was performed for each user (defined as the number of patterns of the user with fastest completion time, which was approximately 6 minutes in the HiMotion dataset and 3 minutes in the PTB diagnostics dataset).

For performance evaluation of the proposed methods, feature extraction data (in the fiducial approach), and strings (in the non-fiducial approach), from both the HiMotion ECG database and the PTB diagnostic ECG database were considered. Results are computed and averaged over 50 runs, in which the complete database is divided into smaller subsets with the following distribution: 70% of the patterns are used as test set for performance evaluation; the remaining 30% of the patterns are used as training set.

In terms of the time required for recognition, we were able to test each method in different conditions of the training and test sets. With the HiMotion database, we used 20 templates for training, and 8 templates for testing, which given our mean waveforms approach, correspond respectively to 200 and 80 seconds of acquired signals. However, for the PTB database, only 4 templates for training and 2 for testing were used, which means an improvement towards 40 seconds of training data and 2 seconds of test data.

D.6.2 Methodology

For the fiducial method, based on each of the mean waves in the database, a new database was built with the features extracted using the process previously described in Section D.5.1. Regarding the non-fiducial method, based on each of the given databases, a new database was built with the corresponding strings for each subject, using the quantization methods described in Section D.5.2.1.

In each of the 50 runs, we make a random partition of the database, into two subsets: one subset containing n patterns is used to form the training set $\mathcal{X}$; the other subset contains the remaining patterns for that subject, which are grouped into test patterns of size m, and used as a test set. Thus, we are using a leave-m-out cross validation strategy.

The authentication performance is assessed using a ROC-based method to determine the *equal error rate* (EER). The class assignment for *false acceptance rate* (FAR) and *false rejection rate* (FRR) calculation was performed based on an user tuned threshold over the shortest distance between a given unknown pattern $\mathbf{z}$ and each known template $\mathbf{x}_u$ in the training set.

An intra-class study with the HiMotion database [27], in the context of the use of electrophysiological signals for emotional states detection, showed the existence of differentiated states in the

ECG data of a subject; to deal with these intra-class differences, the proposed methods include in the "model", single heartbeats randomly selected from the whole set of subject patterns. The same procedure was adopted when using the PTB diagnostic ECG database.

D.6.3 Results

The achieved results are summarized in Table D.2 for both proposed approaches over the considered databases, HiMotion ECG database and the PTB diagnostic ECG database.

Table D.2: Performance comparison of the proposed approaches. The accuracy values (and standard deviation – std) refer to person identification, while the EER values refer to authentication; Subj. is the number of subjects in the database, where the HiMotion ECG database and the PTB diagnostic ECG database were considered.

Database	**Subj.**	**Approach**	**Identification accuracy (std)**	**Authentication EER (std)**
HiMotion	26	Fiducial	99.57% (0.29%)	0.70% (0.15%)
HiMotion	26	Non-fiducial	99.94% (0.24%)	0.29% (0.95%)
PTB	51	Fiducial	99.85% (0.41%)	0.01% (0.02%)
PTB	51	Non-fiducial	99.39% (0.89%)	0.13% (0.42%)

D.6.3.1 Identification Results

For the proposed fiducial method, human identification results were evaluated using the previously described methodology. For the PTB database, we obtained an average recognition rate of 99.85% with a standard deviation (std) of 0.41%, while for the HiMotion database an average recognition rate of 99.57% with a std of 0.29% was attained.

Regarding the non-fiducial method for identification, experiments were done using uniform quantization according to the previously described methodology. Several combinations of the number m of test patterns and the number n of "model" patterns were evaluated to determine the best performance.

The results obtained with the HiMotion database are depicted in Figure D.8, which illustrates the evolution of the average classification error by varying the numbers m and n. Regarding the PTB database, the results are shown in Table D.3. We can observe that the accuracy increases both with n and m, as expected, and the proposed method achieves 99.94% with a std of 0.24% accuracy, for $m = 8$ and $n = 20$, over the HiMotion database, while over the PTB database with $m = 2$ and $n = 4$ (given that less data per subject is available) an accuracy of 99.39% with a std of 0.89% is achived.

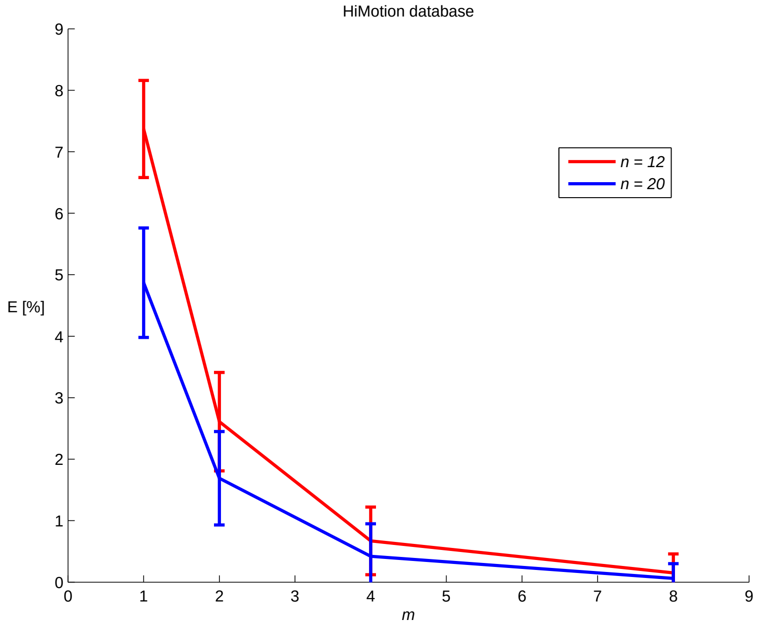


Figure D.8: Non-fiducial ECG-based human identification classification error for the 26 subjects from the the Himotion database. Error results varying both the number of patterns (single heartbeat waveforms) used for the "model" and the testing set. *m*: number of patterns used for the testing set; *E*: average classification error and standard deviation bars. *n*: number of patterns used for the model.

D.6.3.2 Authentication Results

Concerning authentication, the performance evaluation of the proposed fiducial approach for the PTB database presented an average EER of 0.01% with a standard deviation (std) of 0.02%, while for the HiMotion database an average EER of 0.70% with a std of 0.15% was attained. Figure D.9 shows the receiver operating characteristic (ROC) curves evaluated with the HiMotion database for both the fiducial and non-fiducial approaches.

The non-fiducial method was tested using both uniform and non-uniform quantization, given that we wanted to study the effect of user tuned quantization on the system performance. This was evaluated for the HiMotion database and for the PTB database, and Table D.4 shows the best results for the authentication experiments. The values shown are the average EER and the standard deviation (std) over all users from the tested sets, and were obtained with $m = 8$ and $n = 20$ for the HiMotion database, and with $m = 2$ and $n = 4$ for the PTB database. These results show that Lloyd-Max (rather than uniform) quantization and user-adjusted thresholds clearly improves the performance.

Table D.3: Non-fiducial ECG-based human identification classification error for the 51 subjects from the PTB database. Error results for a single number of patterns (single heartbeat waveforms) used for the "model" (*n=4*) and for two different number of patterns m from the testing set (*m*). The values shown are all in percentage and the number in parentheses denotes the standard deviation.

PTB database	$m = 1$	$m = 2$
average identification accuracy	98.81% (0.85%)	99.39% (0.89%)

Table D.4: Comparison of non-fiducial ECG-based authentication results for the 26 subjects from the HiMotion database and the 51 subjects from the PTB database. The presented values are the average EER and the standard deviation (std) over all users from both the databases, considering that the number of patterns (single heartbeat waveforms) used for the "model" is $n = 20$ (HiMotion) or $n = 4$ (PTB), and that the number of patterns used from the testing set is $m = 8$ (HiMotion) or $m = 2$ (PTB).

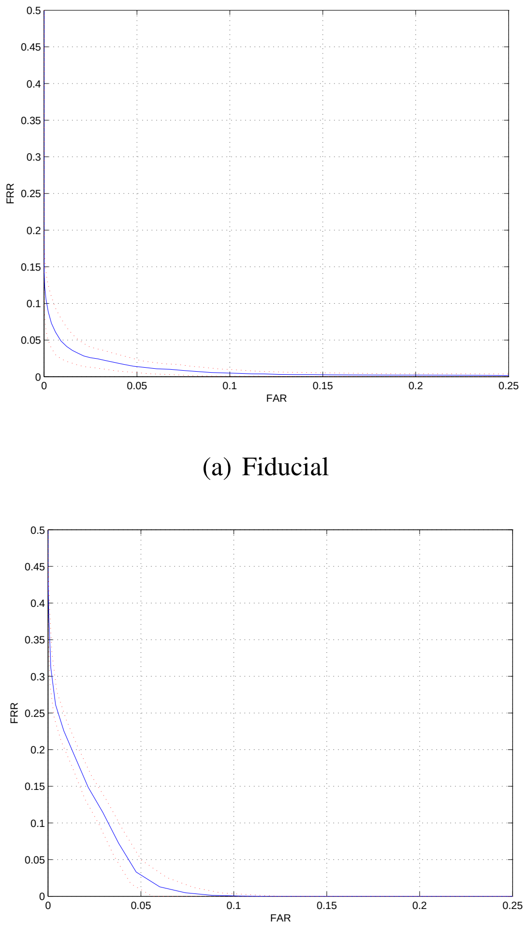
Feature	**HiMotion EER (std)**	**PTB EER (std)**
uniform quantization	$\approx 6\%$	-
uniform quantiz., user-tuned	0.33% (0.88%)	0.22% (0.63%)
Lloyd-Max quantiz., user-tuned	0.29% (0.95%)	0.13% (0.42%)

D.6.3.3 Time Covariate

The time covariate, or permanence of the biometric template [21] is an important property for applications where modalities are used in a standalone format; as a foreword, it's important to highlight that this is not necessarily the case for the ECG, since it's intrinsic properties are already highly valuable in the modern trend towards multibiometric systems [32]. We can easily envision scenarios where a first validation is performed by acquiring simultaneously ECG data with a hard biometrics (e.g. the fingerprint), and the ECG is used afterwards to continue the identity validation process for a period of several hours or days.

Although the permanence of the ECG signal in the context of biometric recognition has not yet been extensively studied, there are at least two authors that have performed preliminary research work in this topic. In the paper by Wübbeler et al. [43], the authors have performed a feasibility study involving 74 subjects from which data was repetitively collected over a period of several months and even years. The average time between recordings was reported to be 16.6 months, and in these conditions, the authors show encouraging results, with 2.8% EER for authentication, and 98.1% accuracy for identification.

The work by Agrafioti et al. [1], motivated by the need to have time invariant templates, proposed



(a) Fiducial

(b) Non-Fiducial

Figure D.9: ROC curves for both Fiducial and Non-Fiducial approaches evaluated with the HiMotion dataset.

a method for template updating over time. The authors supported their study on data collected for a total of 10 subjects within a period of 2 hours, which does not represent a very large time gap between the recognition moments; still, the importance of their work is that it highlights that even though templates may change over time, there are solutions to overcome this, namely by updating the template database.

More recently, Agrafioti et al. [2] [16] have further extended their work by analyzing data collected in two distinct moments in time, separated by a 1-month interval; the study involved 16 subjects, and the authors analyzed the performance in an authentication task, achieving a 14% EER best-case scenario. In the overall, as highlighted by [43], the permanence of the ECG biometric template requires further research, namely in terms of longitudinal studies involving a high number of subjects.

Although not yet fully consistent, the existing references found in literature show that on one hand, there are promising results pointing towards the permanence of the ECG signal over considerably large periods of time [43], and on the other hand, that the system can be designed to cope with the changing templates [1]. Our work builds on these baselines, to explore other dimensions of the ECG biometrics problem.

D.7 Conclusions

We presented and evaluated two biometric identification and authentication techniques based on the electrical activity of the heart (ECG). We addressed the problem from a practical perspective, by considering an unconstrained acquisition scenario and the least amount of data necessary to accurately identify or authenticate a human subject. The proposed approaches are targeted at real-time operation and require minimal data for subject identification and authentication.

For the fiducial approach, results have shown that, from a single mean waveform pattern, we were able to obtain a 99.57% identification accuracy and an EER of 0.70% in authentication, over 26 subjects from the HiMotion database, while for 51 subjects from the PTB database, we obtained a 99.85% recognition rate and an EER of 0.01% in authentication.

The non-fiducial approach results have shown that more than a single mean waveform pattern is needed as test sample to obtain similar results. Using 8 mean waveform patterns, we were able to obtain 99.94% identification accuracy and an EER of 0.29% in authentication, over the same 26 subject's from the database; with the same 51 subjects from the PTB database, and using 2 mean waveform patterns, we obtained a 99.39% identification rate and an EER of 0.13% in authentication.

We can conclude that the fiducial approach achieves very good performance with less data, *i.e.*, with a single mean heartbeat waveform used for test pattern. Nevertheless, the non-fiducial approach has the advantage of not requiring feature extraction, thus not relying critically on the detection of some fiducial points within the ECG signal. The final choice for one of the methods will depend on the system designer, given the application requirements and constraints.

Future work regarding the non-fiducial approach must include further tests in order to study the problem of how to find the optimal number of model (training) and testing samples to be selected. Another challenging issue to be addressed is the permanence (time covariate) study and evaluation for the proposed methods.

References

[1] Foteini Agrafioti, Francis M. Bui, and Dimitrios Hatzinakos. Medical biometrics: the perils of ignoring time dependency. In *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, 2009. BTAS'09*, pages 1–6, 2009.

[2] Foteini Agrafioti, Jiexin Gao, and Dimitrios Hatzinakos. Heart Biometrics: Theory, Methods and Applications, Biometrics. In *Biometrics: Book 3*, pages 199–216. InTech, 2011.

[3] A. Azzini and S. Marrara. Impostor Users Discovery Using a Multimodal Biometric Continuous Authentication Fuzzy System. *Lecture Notes in Computer Science*, 5178:371–378, 2008.

[4] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. Language Trees and Zipping. *Physical Review Letters*, 88(4):048702, January 2002.

[5] L. Biel, O. Pettersson, L. Philipson, and P. Wide. ECG analysis: a new approach in human identification. *IEEE Transactions on Instrumentation and Measurement*, 50(3):808–812, June 2001.

[6] Nikolaos V. Boulgouris, Konstantinos N. Plataniotis, and Evangelia Micheli-Tzanakou. *Biometrics: theory, methods, and applications*. Wiley-IEEE Press, 2009.

[7] Daniele Cerra and Mihai Datcu. Algorithmic Cross-Complexity and Relative Complexity. *2009 Data Compression Conference*, pages 342–351, March 2009.

[8] ADC Chan and MM Hamdy. Wavelet distance measure for person identification using electrocardiograms. *Instrumentation and Measurement, IEEE Transactions on*, 57(2):248–253, 2008.

[9] Chuang-Chien Chiu, Chou-Min Chuang, and Chih-Yu Hsu. A Novel Personal Identity Verification Approach Using a Discrete Wavelet Transform of the ECG Signal. In *2008 International Conference on Multimedia and Ubiquitous Engineering (MUE 2008)*, pages 201–206. IEEE, 2008.

[10] E. Chung. *Pocketguide to ECG Diagnosis*. Blackwell Publishing, 2000.

[11] David Pereira Coutinho and Mário A. T. Figueiredo. Information Theoretic Text Classification Using the Ziv-Merhav Method. In *Pattern Recognition and Image Analysis. Springer Berlin Heidelberg*, volume 1, pages 355–362, 2005.

[12] David Pereira Coutinho, Ana L.N. Fred, and Mário A.T. Figueiredo. One-Lead ECG-based Personal Identification Using Ziv-Merhav Cross Parsing. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010)*, pages 3858–3861. IEEE, August 2010.

[13] J. Cunha, Bernardo Cunha, William Xavier, Nuno Ferreira, and A. Pereira. Vital-Jacket: A wearable wireless vital signs monitor for patients mobility. In *Proceedings of the Avantex Symposium*, 2007.

[14] I. K. Duskalov, Ivan A. Dotsinsky, and Ivailo I. Christov. Developments in ECG acquisition, preprocessing, parameter measurement, and recording. *IEEE Engineering in Medicine and Biology Magazine*, 17(2):50–58, 1998.

[15] H. Gamboa, H. Silva, and A. Fred. HiMotion Project. Technical report, 20070731, IT - Instituto de Telecomunicações, 2007.

[16] Jiexin Gao, Foteini Agrafioti, Hoda Mohammadzade, and Dimitrios Hatzinakos. ECG for blind identity verification in distributed systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1916–1919, 2011.

[17] Z. Geradts and A. Ruifrok. Extracting Forensic Evidence from Biometric Devices. *Proceedings of SPIE*, 5108:181–188, 2003.

[18] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation, Am. Heart Assoc.*, 101(23):e215–e220, 2000.

[19] Fredrik Gustafsson. Determining the initial states in forward-backward filtering. *IEEE Transactions on Signal Processing*, 44(4):988–992, 1996.

[20] Steven A. Israel, John M. Irvine, Andrew Cheng, Mark D. Wiederhold, and Brenda K. Wiederhold. ECG to identify individuals. *Pattern Recognition*, 38(1):133–142, January 2005.

[21] A. Jain, P. Flynn, and A. Ross. *Handbook of Biometrics*. Springer, 2007.

[22] U. Kunzmann, G. Wagner, J. Schöchlin, and A. Bolz. Parameter extraction of ECG signals in real-time. *Biomedizinische Technik/Biomedical Engineering*, 47(s1b):875–878, 2002.

[23] G. Kwang, R. Yap, T. Sim, and R. Ramnath. An Usability Study of Continuous Biometrics Authentication. *Lecture Notes in Computer Science*, 5558:828–837, 2009.

[24] Vladimir Leonov, Tom Torfs, Inge Doms, Refet Firat Yazicioglu, Ziyang Wang, Chris Van Hoof, and Ruud Vullers. Wireless body-powered electrocardiography shirt. In *Proceedings of the 3rd European Conference on Smart Systems Integration*, pages 307–314, 2009.

[25] R. Lourenço, P. Leite, A. Lourenço, H. Silva, A. Fred, and D. P. Coutinho. Experimental Apparatus for Finger ECG Biometrics. In *Proceedings of the International Conference on Biomedical Electronics and Devices (BIODEVICES 2012)*, pages 196–200, 2012.

[26] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino. Impact of Artificial “Gummy” Fingers on Fingerprint Systems. *Proceedings of SPIE*, 4677:275–289, 2002.

[27] Liliana Medina and Ana Fred. Genetic Algorithm for Clustering Temporal Data-Application to the Detection of Stress from ECG Signals. In *Proceedings of 2nd International Conference on Agents and Artificial Intelligence (ICAART)*, pages 135–142, 2010.

[28] K. Niinuma and A. K. Jain. Continuous user authentication using temporal information. *Proceedings of SPIE*, 7667:76670L–76670L, 2010.

[29] Julien Penders, Bert Gyselinckx, Ruud Vullers, Olivier Rousseaux, Mladen Berekovic, Michael Nil, Chris Hoof, Julien Ryckaert, RefetFirat Yazicioglu, Paolo Fiorini, and Vladimir Leonov. Human++: Emerging technology for body area networks. In *VLSI-SoC: Research Trends in VLSI and Systems on Chip*, pages 377–397. Springer US, 2007.

[30] Alejandro Riera, Stephen Dunne, Ivan Cester, and Giulio Ruffini. Starfast: a wire-less wearable EEG/ECG biometric system based on the enobio sensor. In *Proceedings of the International Workshop on Wearable Micro and Nanosystems for Personalised Health*, 2008.

[31] M. Rodriguez, A. Ayala, S. Rodriguez, F. Rosa, and Mario Diaz-Gonzalez. Application of the MaxLloyd quantizer for ECG compression in diving mammals. *Computer Methods and Programs in Biomedicine*, 73(1):13–21, January 2004.

[32] Arun Abraham Ross, Karthik Nandakumar, and Anil K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, 2006.

[33] T. Shen. *Biometric Identity Verification Based on Electrocardiogram*. Phd thesis, University of Wisconsin, 2005.

[34] T. W. Shen, W. J. Tompkins, and Y. H. Hu. One-lead ECG for identity verification. In *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, pages 62–63, 2002.

[35] S. Shepherd. Continuous authentication by analysis of keyboard typing characteristics. In *Proceedings of the European Convention on Security and Detection*, pages 111–114, 1995.

[36] H. Silva, H. Gamboa, V. Viegas, and A. Fred. Wireless Physiologic Data Acquisition Platform. In *Proceedings of the 2005 Conference on Telecommunications*, 2005.

[37] H. Silva, A. Lourenço, R. Lourenç, P. Leite, D. P. Coutinho, and A. Fred. Study and evaluation of a single differential sensor design based on electro-textile electrodes for ECG biometrics applications. In *Proceedings of the IEEE Sensors Conference*, pages 1764–1767, 2011.

[38] Hugo Silva, Hugo Gamboa, and Ana Fred. Applicability of lead v2 ECG measurements in biometrics. In *Proceedings of Med-e-Tel*, 2007.

[39] H. H. So and K. L. Chan. Development of QRS detection method for real-time ambulatory cardiac monitor. In *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, volume 289, pages 289–292, 1997.

[40] S. Suppappola and Y. Sun. A Comparison of Three QRS Detection Algorithms Using the AHA ECG Database. *IEEE Engineering in Medicine and Biology Society*, 13:586–587, 1991.

[41] Yongjin Wang, Foteini Agrafioti, Dimitrios Hatzinakos, and Konstantinos N. Plataniotis. Analysis of Human Electrocardiogram for Biometric Recognition. *EURASIP Journal on Advances in Signal Processing*, 2008(1):19, 2008.

[42] T. Wrublewski, Y. Sun, and J. Beyer. Real-time Early Detection of R Waves of the ECG Signals. *IEEE Engineering in Medicine and Biology Society*, 1:38–39, 1989.

[43] Gerd Wübbeler, Manuel Stavridis, Dieter Kreiseler, Ralf-Dieter Bousseljot, and Clemens Elster. Verification of humans using the electrocardiogram. *Pattern Recognition Letters*, 28(10):1172–1175, 2007.

[44] J. Ziv and a. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977.

[45] J Ziv and N Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39(4):1270–1279, 1993.

Appendix E

An Information Theoretic Approach to Text Sentiment Analysis

Paper E

An Information Theoretic Approach to Text Sentiment Analysis

David Pereira Coutinho and Mário Figueiredo

In *Proceedings of 3rd International Conference on Pattern Recognition Applications and Methods - ICPRAM 2013*, pages 577–580

Abstract

Most approaches to text sentiment analysis rely on human generated lexicon-based feature selection methods, supervised vector-based learning methods, and other solutions that seek to capture sentiment information. Most of these methods, in order to yield acceptable accuracy, require a complex preprocessing stage and careful feature engineering. This paper introduces a coding-theoretic-based sentiment analysis method that dispenses with any text preprocessing or explicit feature engineering, but still achieves state-of-the-art accuracy. By applying the Ziv-Merhav method to estimate the relative entropy (Kullback-Leibler divergence) and the cross parsing length from pairs of sequences of text symbols, we get information theoretic measures that make very few assumptions about the models which are assumed to have generated the sequences. Using these measures, we follow a dissimilarity space approach, on which we apply a standard support vector machine classifier. Experimental evaluation of the proposed approach on a text sentiment analysis problem (more specifically, movie reviews sentiment polarity classification) reveals that it approximates the state-of-the-art, despite being much simpler than the competing methods.

E.1 Introduction

The task of automatically classifying a text, not in terms of topic, but according to the overall sentiment it expresses, is the objective of text *sentiment analysis* (SA). A particular instance of this task is that of determining whether a user review (*e.g.*, of a movie, or a book) is positive or negative, that is, determining the so-called *sentiment polarity*. To solve this binary categorization problem, different approaches have been proposed in the literature. Most of those approaches rely on human-generated lexicon-based feature selection methods, based on which it is possible to build supervised vector-based learning methods. The key drawback of those methods is that they demand a complex preprocessing stage and can only achieve acceptable accuracy with careful lexicon and feature design/engineering.

In this paper, we propose a new information-theoretic approach to text sentiment analysis, and illustrate it in the particular case of binary sentiment polarity categorization. The proposed method does not use any of the classical text preprocessing steps, such as stop-word removal or stemming. The proposed method follows earlier work [1] in that it is based on the *Ziv-Merhav method* (ZMM) for the estimation of relative entropies (or Kullback-Leibler divergences) between pairs of sequences of text symbols, with these estimates serving as features, based on which a classifier (*e.g.*, a support vector machine – SVM) can be built.

The seminal work on the text sentiment analysis problem was published in 2002 by Pang and Lee [8], who focused on movie review sentiment polarity categorization. The method proposed by those authors is based on a human-generated lexicon, based on which bag-of-words (BoW) descriptions of the texts were obtained and used as feature vectors by an SVM classifier. Due to the success of Joachims [3] in dealing with text classification problems by combining SVM classifiers with BoW-based vector space models, many researchers have followed similar approaches.

In this work, we aim at dispensing with the human-generated lexicon for building BoW features, or the need for any other feature design or engineering. For that purpose, we partially follow previous work [1] in that we use the ZMM as a *model-free* feature extractor that doesn't require any human intervention. We adopt the dissimilarity space approach (see [9], [10], and references therein); in particular, we characterize each text by the vector of its ZMM-based dissimilarity values with respect to (all or a subset of the) other texts in the training set. Finally, a standard SVM is used as a classifier. We stress again that the crucial aspect of the proposed approach is that it dispenses with any preprocessing (such as stop-word removal and word stemming) or any human-based feature design. Still, as shown in the experiments reported below, our approach establishes a new state-of-the-art accuracy in a benchmark movie review sentiment polarity categorization dataset.

The outline of the paper is has follows. Section 2 discusses some previous work and results in text sentiment analysis. Section 3 introduces the fundamental tools used in our approach and provides details about our categorization method. Our experiments and analysis of the results are presented in section 4, and finally conclusions are presented in Section 5.

E.2 Related Work

Starting with the seminal work of Joachims [3], SVM classifiers have been one of the weapons of choice when dealing with topic-based text classification. These SVM classifiers typically work on vector spaces where each text is characterized by a bag of words (BoW) or bag of pairs of words (word bi-grams). It was thus not surprising that the initial attempts at addressing text sentiment analysis (which of course is just a special type of text categorization) were also based on SVM tools and BoW-type features [8]. The early work of Pang and Lee, using this type of approach, provided a strong baseline accuracy of 82.9% in a task of movie reviews sentiment polarity (binary) classification.

Since then, the movie review dataset (also known as the sentiment polarity dataset) used in [8], [7] has become a benchmark for many sentiment classification studies. We now recall some of the best result to date on this dataset.

Whitelaw and collaborators [14], reported an accuracy of 90.2%. Their method is based on so-called *appraisal groups*, which are defined as coherent groups of words around adjectives that together express a particular opinion, such as “very funny” or “not terribly surprising”. After building an apraisal lexicon (manually verified) it uses a combination of different types of appraisal group features and BoW features for training an SVM classifier. The state-of-the-art accuracy was established by Matsumoto and collaborators [6]. They proposed a method where information about word order and syntactic relations between words in a sentence is used for training a classifier. Thus using the extracted word sub-sequences and dependency sub-trees as features for SVMs training they attained an accuracy of 93.7%.

More recently Yessenalina and colleagues [15] proposed a supervised multi-level structured model based on SVMs, which learns to jointly predict the document label and the labels of a sentence subset that best explain the document sentiment. The authors treated the sentence-level labels as hidden variables so the proposed model does not required sentence-level annotation for training, avoiding this way the lowerlevel labellings cost. They formulate the training objective to directly optimize the document-level accuracy. This multi-level structured model achieved 93.22% document-level sentiment classification accuracy on the movie review dataset.

These results and references are summarized in Table 2.3. Further examples can be found in the survey paper [13], but all usually involving complex preprocessing stages and careful feature engineering.

Table E.1: Baseline and best reported classification accuracies in the literature over the same collection of movie reviews.

Method	Accuracy [%]
(Pang et al., 2002) [8]	82.9
(Pang and Lee, 2004) [7]	87.2
(Whitelaw et al., 2005) [14]	90.2
(Matsumoto et al., 2005) [6]	93.7
(Kennedy and Inkpen, 2006) [4]	86,2
(Yessenalina et al., 2010) [15]	93.2
(Maas et al., 2011) [5]	88,9
(Duric and Song, 2011) [2]	87,5
Proposed approach with linear SVM	87.0

E.3 Proposed approach

E.3.1 The Ziv-Merhav Method

The Ziv-Merhav Method (ZMM) was introduced in 1993 [18] for measuring relative entropy between pairs of sequences of symbols. It is based on the incremental Lempel-Ziv (LZ) parsing algorithm [17] and on a variation thereof, known as cross parsing. Combining these two algorithms, the authors defined an estimate of the relative entropy that can be used as a dissimilarity measure.
The LZ algorithm is a well-known tool for text compression [12], which in recent years has also been used for text/sequence classification purposes; for example, in [1], LZ-based dissimilarity measures were used to achieve state-of-the-art performance in a specific text classification task (authorship attribution).
An implementation of the cross parsing algorithm was proposed in [1], based on a modified LZ77 [16] algorithm, where the dictionary is static and only the lookahead buffer slides over the input sequence, as shown in Figure 3.2 (for more details, see [1]). This very same implementation, using a 2 Mbyte dictionary and a 256 byte look ahead buffer, was used in the experiments reported below.

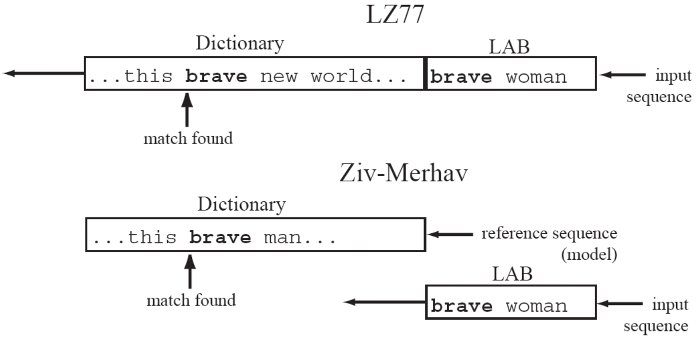


Figure E.1: The original LZ77 sliding window and the modified implementation for cross parsing.

Whereas in [1], the ZMM was applied to compute text dissimilarities, which were then used by a K-nearest-neighbors (K-NN) classifier, here we propose to use the ZMM to build a dissimilarity space representation of the texts, following the framework proposed in [9], [10], and reviewed in the next subsection.

E.3.2 Dissimilarity-Based Classification

Let us consider a given training set of objects (movie reviews, in the example considered in this paper) $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, where each object belongs to some set $\mathcal{X}$ (*e.g.*, the set of finite length strings of some finite alphabet) and some dissimilarity measure between pairs of objects, $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In the dissimilarity-based approach, each object (either in the training set or a new object to be classified after training) is represented by the vector of its dissimilarities with respect to the elements of $\mathbf{X}$ (or a subset thereof). That is, the training set in the so-called dissimilarity space becomes

$$\mathcal{D} = \{\mathbf{d}_1, ..., \mathbf{d}_n\},$$

where

$$\mathbf{d}_i = \begin{bmatrix} D(\mathbf{x}_i, \mathbf{x}_1) \\ \vdots \\ D(\mathbf{x}_i, \mathbf{x}_n) \end{bmatrix} \in \mathbb{R}^n.$$

An important aspect of dissimilarity-based approaches is that very few conditions are put of the dissimilarity measure; namely, it doesn't have to be a metric, it doesn't even need to be symmetric [9], [10]. Dissimilarity representations can also be based on a subset of the training set, in which case the dissimilarity space has dimension equal to the cardinal of that subset; some work has been devoted to methods for selecting this subset [11].

In this paper, we propose to build the dissimilarity-based representation by using the relative entropy estimate between pairs of sequences of symbols, as measured by the Ziv-Merhav method described in the previous subsection.

Once in possession of a dissimilarity-based representation of a training set, any standard classification method that works on vector spaces can be used. In this paper, we report preliminary results by using (linear) SVM and K-NN classifiers. As shown in the experiment results below, this simple approach already achieves results that outperform the previous state-of-the-art, although it is conceptually much simpler and requires much less human intervention. In future work, even better results may be obtained by exploring other possibilities, such as other kernels, tuning of the SVM C parameter, and better strategies to select a subset of objects with respect to which the dissimilarity representations are obtained.

E.4 Experimental Setup

In the experimental evaluation of the proposed approach, we use the polarity dataset[1] v2.0, introduced by [7]; this (human classified) dataset includes 1,000 positive and 1,000 negative movie reviews. The dataset is split into a training set with 900 examples per class and then into 10 cross-validation (CV) folds. We report CV accuracy estimates, following the same protocol of [7], where in each run, 1800 examples are used to train and 200 examples to test. We stress, that we don't use any text preprocessing.

We use K-NN and SVM classifiers (with linear kernel), implemented by the PRTools Matlab toolbox for pattern recognition [2] (version 4). The value of the C parameter in the SVM was set to one.

[1] www.cs.cornell.edu/people/pabo/movie-review-data

[2] www.prtools.org/index.html

E.5 Results

We compare the accuracy of the proposed approach with respect to the methods described in Section E.2 in the movie review sentiment polarity classification problem using the dataset described in the previous section. The results shown in Table 2.3 reveal that the proposed approach with the SVM approximates state-of-the-art by achieving an average accuracy of 87.0%, outperforming Pang and Lee's first results on this dataset [8], although is drastically simpler and requires much less human intervention. Regarding the K-NN classifier, experimental results are much worse than the SVM results.

Finally, we also explored the random prototype selection method proposed by Duin et al. [11]; the results are shown in Figure E.2. Notice that using only 30% of the prototypes for training (540) is almost as better than Pang and Lee's first result.

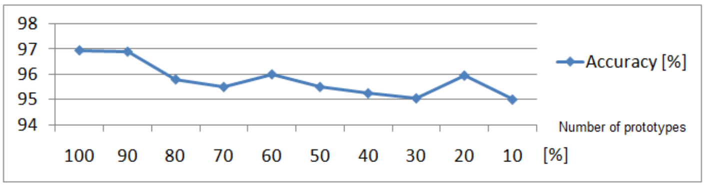


Figure E.2: Average accuracy in the dissimilarity space as a function of the number of randomly prototypes.

E.6 Conclusions

In this paper, we have presented a new approach for text sentiment analysis, based on an information-theoretic dissimilarity measure, which is used to build dissimilarity representations on which SVM and K-NN classifiers were applied. The aim of our proposal was mainly to show that this type of approach approximates the state-of-the-art results in hard text classification problems, while involving virtually no human intervention and no text preprocessing. We have illustrated the approach on a benchmark dataset, where the task is to perform movie review sentiment polarity categorization. Experimental results showed that K-NN classifiers doesn't work well but the proposed method using SVM classifiers approximates state-of-the-art results, although it is drastically simpler and requires much less human intervention. Future work will include further experiments on other (larger) datasets to assess the relative performance, exploring other kernels and fine tuning of the SVM C parameter and search for better strategies to select a subset of objects (models) with respect to which the dissimilarity representations are obtained.

References

[1] David Pereira Coutinho and Mário A. T. Figueiredo. Information Theoretic Text Classification Using the Ziv-Merhav Method. In *Pattern Recognition and Image Analysis. Springer Berlin*

Heidelberg, volume 1, pages 355–362, 2005.

[2] Adnan Duric and Fei Song. Feature selection for sentiment analysis based on content and syntax models. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 96–103, 2011.

[3] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, 1998.

[4] Alistair Kennedy and Diana Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22(2):110–125, May 2006.

[5] A. L. Maas, R. E. Daly, P. T. Pham, and Dan Huang. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics*, pages 142–150, 2011.

[6] Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 301–311, 2005.

[7] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics*, page 271, 2004.

[8] Bo Pang, Lillian Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86, 2002.

[9] E. Pekalska, P. Paclik, and R. P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.

[10] Elbieta Pkalska and Robert P. W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, 23(8):943–956, June 2002.

[11] Elbieta Pkalska, Robert P. W. Duin, and Pavel Paclík. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, February 2006.

[12] David Salomon and Giovanni Motta. *Handbook of Data Compression*. Springer, 2010.

[13] G. Vinodhini and R. M. Chandrasekaran. Sentiment Analysis and Opinion Mining: A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6), 2012.

[14] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using Appraisal Taxonomies for Sentiment Analysis. In *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*, 2005.

[15] Ainur Yessenalina, Y. Yue, and C. Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics*, pages 1046–1056, 2010.

[16] J. Ziv and a. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977.

[17] J. Ziv and a. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, September 1978.

[18] J Ziv and N Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39(4):1270–1279, 1993.

.

Appendix F

On Compression-Based Text Authorship Attribution

Paper F

On Compression-Based Text Authorship Attribution

David Pereira Coutinho and Mário Figueiredo

In *Proceedings of the 19th edition of the*
Portuguese Conference on Pattern Recognition - RECPAD 2013, number 4

Abstract

Choosing an appropriate set of features that allows a machine learning algorithm to accurately solve a given problem is arguably one of the most difficult tasks in text classification. Most state-of-the-art approaches involve careful feature engineering following a preprocessing stage, which may be too expensive in the emerging context of massive collections of electronic texts. In this paper, we propose efficient methods for text classification, based on information-theoretic dissimilarity measures, which are used to define dissimilarity-based representations. These methods dispense with any feature design or engineering, by mapping texts into a feature space using universal *dissimilarity measures; in this space, classical classifiers (*e.g. *nearest neighbor or support vector machines) can then be used. The reported experimental evaluation of the proposed methods, on a benchmark authorship attribution problem, reveals that it outperforms previous methods, despite being much simpler, in the sense that it does not require any pre-processing or feature engineering.*

F.1 Introduction

Text classification (or categorization) is the problem of assigning a text to one or more of a predefined set of classes. Examples of applications are: (i) *topic classification*, where the task is to decide which topic(s) is (are) addressed in a text; (ii) *sentiment analysis*, which is the task of automatically classifying a text, not in terms of topic, but according to the overall sentiment it expresses; (iii) *authorship attribution* (AA), where the task is to assign a text of an unknown author to one of a set of possible authors. Classical techniques for these (and other) text classification tasks are based on statistical and computational tools that require careful feature engineering and sophisticated preprocessing, which may become prohibitive in the emerging context of massive collections of electronic texts, such as product reviews and e-mail messages. Defining a similarity measure between texts (or, more generally, finite sequences of symbols) that allows addressing classification problems, without explicitly modeling their statistical behavior, is a fundamental problem in this context. In this paper, we aim at dispensing with the need for any feature design or engineering. For that purpose, we partially follow our previous work [1], in that we use compression/parsing-based feature extractors that don't require any human intervention or feature design. We adopt a dissimilarity space approach [5], in which each text is characterized by the vector of its (dis)similarities with respect to the other texts in the training set. Finally, standard *nearest-neighbor* (NN) and *support vector machines* (SVM) are used as classifiers. Our experimental results reveal that our approach outperforms previous methods.

F.2 Compression-based dissimilarity measures

In recent years, much work has been done concerning the design and development of information-theoretic dissimilarity measures. Most of the proposed approaches were developed in a Kolmogorov complexity framework (*e.g.* [4]), but an alternative approach can be based on Shannon's relative entropy [6].

F.2.1 The Normalized Compression Distance

One of the best known compression-based dissimilarity measures for text is the *normalized compression distance* (NCD), proposed by Li et al [4]. NCD approximates the (non-computable) Kolmogorov complexity of a string **x** by the length of a compressed version of **x**, using off-the-shelf compressors, such as *gzip* or *bzip2*, and it is defined for any pair of strings **x** and **y** as

$$NCD(\mathbf{x},\mathbf{y}) = \frac{C(\mathbf{x}\circ\mathbf{y}) - \min\{C(\mathbf{x}), C(\mathbf{y})\}}{\max\{C(\mathbf{x}), C(\mathbf{y})\}}, \tag{F.1}$$

where $C(\mathbf{x})$ is the length of string **x** after being compressed by a lossless compression algorithm, and $\mathbf{x}\circ\mathbf{y}$ denotes the concatenation of strings **x** and **y**. The authors demonstrate that it is a metric and claim that it minorizes every computable distance in a certain class. NCD ranges from 0 to $1+\epsilon$, where 0 corresponds to **x** and **y** being identical, and 1 means maximum dissimilarity. The constant ϵ is an upper bound due to imperfections in the compression algorithms, but is unlikely to be above 0.1 for most standard compressors [4].

F.2.2 The Ziv-Merhav Relative Entropy Estimate

A method for estimating the relative (Shannon) entropy between pairs of sequences of symbols was introduced by Ziv and Merhav (ZM) [6] and has been used as a dissimilarity measure for *universal* classification [1]. The ZM method is based on the incremental Lempel-Ziv (LZ) parsing algorithm and on a variation thereof, known as cross-parsing. Combining these two algorithms, Ziv and Merhav proposed an estimator of the relative entropy between two ergodic sources producing the sequences **z** and **x**, which can be used as a dissimilarity measure between those sequences. Specifically, they proved that for two finite order (of any order) Markovian sequences of length n, the quantity

$$\Delta(\mathbf{z}||\mathbf{x}) = \frac{1}{n}\left[\, c(\mathbf{z}|\mathbf{x})\ \log_2 n - c(\mathbf{z})\ \log_2 c(\mathbf{z})\,\right] \tag{F.2}$$

converges, as $n \to \infty$, to the relative entropy between the two sources that emitted the two sequences, where $c(\mathbf{z})$ denotes the number of phrases resulting from the self-parsing of **z** and $c(\mathbf{z}|\mathbf{x})$ is the number of phrases resulting from cross-parsing **z** with respect to **x**. Roughly speaking, we can interpret $(1/n)\, c(\mathbf{z})\ \log_2 c(\mathbf{z})$ as a measure of 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**. The difference between the two quantities thus provides a measure of how different the distributions that produced the two sequences are.

F.2.3 The Cross-Parsing Distance

The use of the Ziv-Merhav relative entropy estimate is not directly applicable in some scenarios, namely because it is defined for sequences of the same length n. When generalizing this definition to sequences of different lengths, several problems arise, with the size of the "model" sequence **x** having a significant impact [3]. To overcome this difficulty, Helmer et al [3] recently introduced the *cross-parsing distance* (CPD), which is a semi-metric (i.e., of all the conditions that have to be satisfied by a metric, it only does not satisfy the triangle inequality) defined for any pair of strings **x** and **y** (of length respectively |**x**| and |**y**|), as

$$\text{dist}_{CPD}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\left(\frac{|s(\mathbf{x}|\mathbf{y})\ \setminus\ \{\,\mathbf{y}\}\,|}{|\mathbf{x}|} + \frac{|s(\mathbf{y}|\mathbf{x})\ \setminus\ \{\,\mathbf{x}\}\,|}{|\mathbf{y}|}\right), \tag{F.3}$$

where $s(\mathbf{x}|\mathbf{y})$ denotes the multiset of all phrases resulting from the cross-parsing of **x** with respect to **y** and $s(\mathbf{x}|\mathbf{y})\ \setminus\ \{\mathbf{y}\}$ denotes the removal of a single instance of **y** from the multiset $s(\mathbf{x}|\mathbf{y})$ (if one exists, even if multiple copies exist). If the first not yet parsed symbol in **x** is not found in **y**, then the parsing is simply the symbol itself (see Helmer et al. [3] for details).

We will use in this paper as a variant of CPD, a modified version of dist_{CPD} which we call ***CPdist***, defined as

$$\text{CP}_{dist}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\left(\frac{c(\mathbf{x}|\mathbf{y})\ -\ 1_{\mathbf{x}=\mathbf{y}}}{|\mathbf{x}|} + \frac{c(\mathbf{y}|\mathbf{x})\ -\ 1_{\mathbf{y}=\mathbf{x}}}{|\mathbf{y}|}\right), \tag{F.4}$$

where $c(\mathbf{x}|\mathbf{y})$ is the number of phrases resulting from the cross-parsing and 1_A is the indicator function of the proposition A.

F.2.4 Incremental Cross-Parsing Algorithm

An implementation of the cross-parsing algorithm based on the LZ77 algorithm (here termed CP77) was proposed by Pereira Coutinho and Figueiredo [1]; this implementation uses a 2 Mbyte dictionary and a 256 byte *look ahead buffer* (LAB), where the dictionary is static and only the LAB slides over the input sequence (for details see [1]). However LZ77 is itself an incremental parsing algorithm;

thus, following the same idea, we propose a new implementation, based on our first implementation but using incremental dictionary updates (termed CP77inc). Now, the cross-parsing of string **z** with respect to the string **x** involves some details, which we briefly describe; it uses one sliding window to hold both the dictionary D_x and the LAB lab_x of the model string **x**. In addition, it uses another (smaller) sliding window to hold the LAB lab_z for the unknown string **z**. Notice that the dictionary D_x is empty at the beginning. Then, a loop is repeated until the end of **z** is reached: the cross-parsing of **z** given **x**; the self parsing of **x** including dictionary update as long as **x** lasts. This makes sequences of different lengths allowed, by stopping the dictionary update whenever the end of **x** is reached and keep using it as a "static" dictionary. Every time the loop is executed, a counter c_{zx} is incremented. Finally, we call the method of relative entropy estimate via definition (F.2) as ***ZMMinc***, which uses the proposed algorithm CP77inc.

F.3 Proposed dissimilarity-based classification

At the core of dissimilarity-based methods for classification is the computation of pairwise dissimilarities between the object (*e.g.* text) to be classified and a set of (or all) objects (*e.g.*, texts) in the training set. Of course, there are several ways to use dissimilarity values to define a classifier, the simplest of which is arguably to use a k-NN classifer; in this case, the object to be classified is simply assigned to the majority class in its k nearest (in the adopted similarity measure) neighbors (with some rule to break ties). A more sophisticated approach is offered by the dissimilarity space approach [5], which uses the dissimilarity values as features that characterize the object to be classified, based on which several different types of classifiers can be used, namely k-NN in the dissimilarity space or support vector machines (SVM).

Let us consider a training collection of objects (texts) $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, where each object belongs to some set $\mathcal{X}$ (*e.g.*, the set of finite length strings of some finite alphabet Σ), and some dissimilarity measure between pairs of objects, $D : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. In the dissimilarity-based approach, each object (either in the training set or a new object to be classified after training) is represented by the vector of its dissimilarities with respect to the elements of $\mathbf{X}$ (or a subset thereof). That is, the training set in the so-called dissimilarity space becomes

$$\mathcal{D} = \{\mathbf{d}_1, ..., \mathbf{d}_n\},$$

where

$$\mathbf{d}_i = \begin{bmatrix} D(\mathbf{x}_i, \mathbf{x}_1) \\ \vdots \\ D(\mathbf{x}_i, \mathbf{x}_n) \end{bmatrix} \in \mathbb{R}^n.$$

In this paper, we propose to use a dissimilarity space approach, where the representations are built by using the dissimilarity/distance measures described in the previous section (see Figure F.1). Once in possession of a dissimilarity-based representation of a training set, any standard classification method can be used.

For example, when using a k-NN classifier, given a new object, its dissimilarity vector $\mathbf{d}_y$ is built and distances in the dissimilarity space are computed (Euclidean distances, for example) between the new vector $\mathbf{d}_y$ and all the vectors in $\mathcal{D}$. Then, the new object is classified in the most common class amongst its k nearest neighbors.

An important aspect of dissimilarity-based approaches is that very few conditions are put of the dissimilarity measure; namely, it doesn't have to be a metric, it doesn't even need to be symmetric [5].

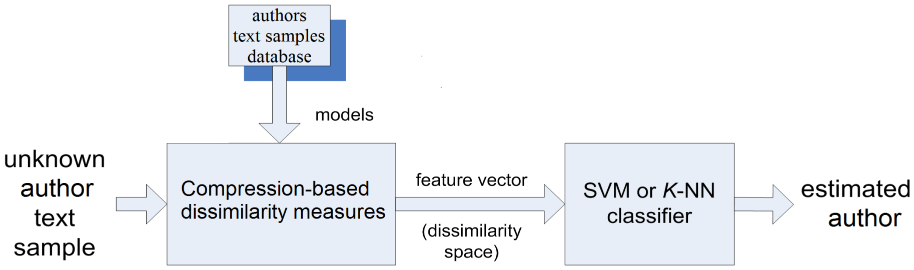


Figure F.1: Proposed system block diagram for text authorship attribution.

F.4 Experiments

Experiments were carried out using both k-NN and (linear) SVM classifiers on the dissimilarity space, with the accuracy assessed by leave-one-out cross-validation (LOO-CV), and using the English Corpus recently introduced by Ebrahimpour et al. [2]. This corpus[1] contains 168 short stories with undisputed authorship by seven English writers of the late 19th century and early 20th century; each story is truncate to approximately the first 5,000 words. Table F.1 shows the accuracy results for that corpus when using NCD, CPdist and ZMMinc as dissimilarity measures. Our best method (ZMMinc and SVM with optimized C parameter), only misclassifies 2 out of the 168 texts.

Table F.1: Leave-one-out cross-validation accuracy percentages, using dissimilarity-based classifiers on a benchmark English Corpus.

		K-NN			SVM		
Corpus	**Baseline** [2]	**NCD**	**CPdist**	**ZMMinc**	**NCD**	**CPdist**	**ZMMinc**
English	96.4	84.5	91.7	87.5	95.2	95.2	**98.8**

[1] Available at `http://promo.net/pg`

F.5 Conclusions

Achieving good accuracy in text classification usually requires careful feature engineering and complex preprocessing stages, which may become prohibitive in the emerging context of classification massive sets of electronic texts. In this paper, we proposed methods for automatic text classification using information-theoretic dissimilarity measures, based on universal data compression algorithms, which bypass the feature design and preprocessing stages.

On an authorship attribution problem, experiments were done using several dissimilarity measures and the best of the proposed methods outperformed the state-of-the-art approaches.

References

[1] David Pereira Coutinho and Mário A. T. Figueiredo. Information Theoretic Text Classification Using the Ziv-Merhav Method. In *Pattern Recognition and Image Analysis. Springer Berlin Heidelberg*, volume 1, pages 355–362, 2005.

[2] Maryam Ebrahimpour, Tlis J. Putniš, Matthew J. Berryman, Andrew Allison, Brian W.-H. Ng, and Derek Abbott. Automated authorship attribution using advanced signal classification techniques. *PloS one*, 8(2):e54998, January 2013.

[3] Sven Helmer, Nikolaus Augsten, and Michael Böhlen. Measuring structural similarity of semistructured data based on information-theoretic approaches. *The VLDB Journal*, 21(5):677–702, February 2012.

[4] M. Li, Xin Chen, and Xin Li. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.

[5] E. Pekalska, P. Paclik, and R. P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.

[6] J Ziv and N Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39(4):1270–1279, 1993.

.

Appendix G

Text Classification Using Compression-based Dissimilarity Measures

Paper G

Text Classification Using Compression-based Dissimilarity Measures

David Pereira Coutinho and Mário Figueiredo

Submitted to the journal of *Pattern Recognition Letters*, 2013

The layout has been revised.

Abstract

Arguably the most difficult task in text classification is to choose an appropriate set of features that allows machine learning algorithms to provide accurate classification. Most state-f-the-art techniques for this task involve careful feature engineering and a preprocessing stage, which may be to expensive in the emerging context of massive collections of electronic texts. In this paper, we propose efficient methods for text classification based on information-theoretic dissimilarity measures, which are used to define dissimilarity-based representations. These methods dispense with any feature design or engineering, by mapping texts into a feature space using universal dissimilarity measures; in this space, classical classifiers (e.g. nearest neighbor or support vector machines) can then be used. The reported experimental evaluation of the proposed methods, on sentiment polarity analysis and authorship attribution problems, reveals that it approximates, sometimes even outperforms previous state-of-the-art techniques, despite being much simpler, in the sense that they do not require any text pre-processing or feature engineering.

G.1 Introduction

Text classification (or categorization) is the problem of assigning a text to one or more of a predefined set of classes. Examples of applications are: (i) *topic classification*, where the task is to decide which topic(s) is (are) addressed in a text; (ii) *sentiment analysis* (SA), which is the task of automatically classifying a text, not in terms of topic, but according to the overall sentiment it expresses, *e.g.*, determining whether a user review of some product or service is positive or negative; (iii) *authorship attribution* (AA), where the task is to assign a text of an unknown author to one of a set of possible authors. Classical techniques for these (and other) text classification tasks are based on statistical and computational tools that require careful feature engineering and sophisticated preprocessing, which may become prohibitive in the emerging context of massive collections of electronic texts, such as product reviews and e-mail messages. Defining a similarity measure between texts (or, more generally, finite sequences of symbols) that allows addressing classification problems, without explicitly modeling their statistical behavior, is a fundamental problem in this context.

In this paper, we aim at dispensing with the need for any feature design or engineering. For that purpose, we partially follow our previous work [6], in that we use compression/parsing-based feature extractors that don't require any human intervention or feature design. We adopt a dissimilarity space approach [22, 23], in which each text is charaterized by the vector of its (dis)similarities with respect to the other texts in the training set. Finally, standard *nearest-neighbor* (NN) and *support vector machines* (SVM) are used as classifiers. Our experimental results reveal that our approach approximates, and in some cases outperforms, state-of-the-art methods for the tasks considered, although, unlike those methods, does not rely on careful pre-processing and feature engineering. In more detail, our contributions are:

- We introduce a vector representation for texts, whose computation dispenses with any preprocessing (such as stop-word removal and word stemming) or any human-based feature design.
- We consider several different compression-based methods to compute these representations, namely the so-called *normalized compression distance* (NCD), the relative entropy estimated via the *Ziv-Merhav method* (ZMM), and the *cross-parsing distance* (CPD).

The outline of the paper is has follows: Section G.2 discusses some previous work and results in text *sentiment analysis* (SA) and *authorship attribution* (AA). Section G.3 reviews some basic concepts from information and complexity theories, while Section G.4 introduces some compression-based dissimilarity measures. Section E.3.2 describes the adopted classification method. Experimental results are presented in Section G.6, and Section G.7 concludes the paper.

G.2 Related Work

Starting with the seminal work of Joachims in 1998 [12], the SVM has been one of the weapons of choice for dealing with (namely topic-based) text classification, working on a vector space where each text is represented by a bag of words (BoW) or bag of pairs of words (word bi-grams). It was thus not surprising that the initial attempts to address text SA (which of course is just a special type of text categorization) were also resorted to SVM tools and BoW-type features [21].

The early work of Pang et al. in 2002 [21], using the type of approach mentioned in the previous paragraph, provided a strong baseline accuracy of 82.9% in a task of movie reviews sentiment polarity (binary) classification. Since then, the movie review dataset introduced by Pang et al. [21] and Pang and Lee [20], also known as the sentiment polarity dataset[1], has become a benchmark in the SA literature. A comprehensive review of recent results on this dataset, some involving sophisticated and carefully engineered features, can be found in the survey papers by Jebaseeli and Kirubakaran [11] and Vinodhini and Chandrasekaran [33].

The Multi-Domain Sentiment Dataset[2], introduced by Blitzer et al. in 2007 [2], is larger and includes four other widely-used datasets with reviews of several different types of products. Recently, Xia et al. [35] used that dataset together with the (movie reviews) sentiment polarity dataset, in a comparative study of the effectiveness of ensembles of several feature sets and classification tools (namely naïve Bayes, maximum entropy, and SVM) for SA; their results can be summarized by the average (over the five datasets) accuracies of 83.12% and 85.58%, for two types of proposed ensemble techniques; we will use these accuracies as the baseline for SA.

The other type of text categorization problem considered in this paper is *authorship attribution* (AA). Addressing AA with statistical and computational tools has a long history, which can arguably be traced back to the seminal study on the authorship of the disputed *Federalist Papers* published by Mosteller and Wallace in 1963 and 1964 [18, 17]. The Federalist Papers are a series of 85 political essays published in 1788 by Alexander Hamilton, James Madison, and John Jay. Initially, the identit y of the author of each essay was kept secret; although later the authors claimed authorship of their essays, some disputes arose. Presently, experts consider that 73 texts can be considered as having known author, while 12 are of disputed authorship. Mosteller and Wallace [17] proposed a Bayesian statistical method based on the frequencies of a small set of common words (*e.g.*, “and”, “to”, ...), which produced good discrimination results. Regarding recent research advances in this field, an extensive survey was presented by Stamatatos in 2009 [30], where the different approaches for text representation and text classification are analyzed, focusing on computational requirements and settings, while discussing evaluation methodologies and criteria for AA problems.

Let us now focus on a particular approach to text classification, based on similarity measures. Benedetto et al. in 2002 [1] proposed a simplified “distance” function between a pair of texts, based on the description length obtained by encoding one text using a code (a model) optimized for the other text; in practice, they propose computing this distance by concatenating the two texts and compressing the result using an off-the-shelf universal encoder, such as *gzip* or *zip*. In order to evaluate the accuracy of their method, Benedetto et al. [1] carried out experiments using a corpus of 90 texts of Italian authors (to which we will refer as the *Italian Corpus*[3], and which we will use as a benchmark dataset), reporting an accuracy of 93.3%. However, that method has some weaknesses. Namely, *gzip* is a dictionary-based compression algorithm that uses a sliding window of length 32 Kbytes to build the dictionary[4]; thus, if the model text is long enough, its beginning will be ignored when *gzip* is compressing its concatenation with the other text. Furthermore, if the other text is long enough, the

[1] Available at `http://www.cs.cornell.edu/people/pabo/movie-review-data`

[2] Available at `http://www.cs.jhu.edu/~mdredze/datasets/sentiment`

[3] Available at `http://www.liberliber.it`

[4] For details see http://www.gzip.org/algorithm.txt

model will disappear from the dictionary after a while. Puglisi et al. [25] studied in detail what happens when a dictionary-based compression algorithm, such as *gzip*, tries to optimize its features at the interface between two different texts.

To overcome some of the weaknesses of the approach of Benedetto et al. [1], Pereira Coutinho and Figueiredo [6] proposed a method based on an estimator of the relative entropy between pairs of sequences of symbols proposed by Ziv and Merhav [38]. An accuracy of 95.4% on the *Italian Corpus* was achieved by a NN classifier based on that relative entropy estimate as a distance measure.

An alternative distance measure based on a computable algorithmic relative complexity was proposed by Cerra and Datcu [3]. On a pre-processed version of the *Italian Corpus*, an accuracy of 97.8% was reported by those authors. This result will be used as the AA baseline result on this dataset.

Finally, very recently, Ebrahimpour et al. [9] introduced a corpus of English texts obtained from the Project Gutenberg archives[5], containing 168 short stories by seven undisputed authors from the late 19th century and early 20th century; this will be referred to as the *English Corpus*. Ebrahimpour et al. [9] claimed that the authors writing styles would be the key discriminant feature, since all the selected authors wrote fictional literature in English of the same genre in the same era; they developed two AA methods, one of which is an SVM with features based on word frequencies, involving some text pre-processing. Experimental results revealed an accuracy of 96.4%. Moreover, the authors applied the same methods to the Federalist Papers and reported an accuracy of 97.1%. These results will be used as AA baseline results for these datasets.

G.3 Preliminaries

G.3.1 Shannon Entropy and Kolmogorov Complexity

The classical information theory has its origin in the work of Shannon [27], who introduced the concept of entropy, which is a measure of the uncertainty about the outcomes of a random variable $X \in \mathcal{X}$ with a given probability mass function $p(x) = P(X = x)$, defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(x)\, \log_2 p(x)\,. \tag{G.1}$$

As is famously known, this quantity can be interpreted as a measure of the average amount of information (expressed in *bits*), needed to describe the random variable X. One *bit* is the amount of information need to describe, on average, the outcome of a uniformly distributed binary random variable (*e.g.*, a fair coin). Moreover, $H(X)$ can be seen as the shortest expected length with which it is possible to encode losslessly the outcomes of X.

The Shannon entropy is an ensemble/average quantity based on probabilistic assumptions; consequently, it does not provide the informational content of individual objects. In contrast, given an instance x (typically a string of symbols) its so-called Kolmogorov complexity $K(x)$, one of the central concepts of algorithmic information theory, is a measure of its intrinsic complexity [16]. Algorithmic information theory has its roots in the seminal work of Ray Solomonoff, Andrey Kolmogorov,

[5] Available at http://www.gutenberg.org

and Gregory Chaitin in the 1960s [28, 29, 13, 4] and includes several quantities defined on strings of symbols, including complexity, randomness, and information.

The Kolmogorov complexity $K(x)$, or algorithmic complexity, of a string x, is a measure of the computational resources needed to describe that string; more precisely, it is the length (usually in bits) of the shortest possible program used as input by a universal Turing machine to produce the string x and halt [16]. It can be shown that the Kolmogorov complexity of any string cannot be more than a few bytes larger than the length of the string itself, while low complexity strings may have considerably shorter program descriptions. One interpretation of $K(x)$ is the quantity of information needed to recover x from scratch. However, it is known that $K(x)$ is non-computable [16], thus approximations must be used, such as the length $C(x)$ of a compressed version of x using some off-the-shelf lossless compressor.

The formal link between Shannon entropy and algorithmic complexity has been established as

$$H(X) \leq \sum_{x} p(x)\, K(x) \ \leq H(X) + K(p) + O(1)\,, \tag{G.2}$$

where $K(p)$ denotes the so-called probability function complexity [16]. For low complexity distributions, the impact of $K(p)$ is lower and the expected Kolmogorov complexity is close to the Shannon entropy.

G.3.2 Kullback-Leibler Divergence and Optimal Coding

Consider two memoryless sources $\mathcal{A}$ and $\mathcal{B}$ producing sequences of binary symbols; source $\mathcal{A}$ emits 0 with probability p (thus 1 with probability $1-p$), while $\mathcal{B}$ emits 0 with probability q. According to Shannon's information theory [27, 8], there are compression algorithms that, applied to sequences emitted by $\mathcal{A}$, are able to encode them with an average number of bits per character asymptotically equal to the source entropy $H(\mathcal{A})$,

$$H(\mathcal{A}) = -p\,\log_2 p - (1-p)\,\log_2(1-p) \quad \text{bits/symbol.} \tag{G.3}$$

An optimal code for $\mathcal{B}$ will not be optimal for $\mathcal{A}$ (unless, of course, $p = q$). The average number of excess bits per symbol that are wasted when we encode sequences emitted by $\mathcal{A}$ using an optimal code for $\mathcal{B}$ is given by the relative entropy, or Kullback-Leibler (KL) divergence, between the corresponding distributions [8], that is

$$D(\mathcal{A}||\mathcal{B}) = p\log_2\frac{p}{q} + (1-p)\log_2\frac{1-p}{1-q}. \tag{G.4}$$

This fact suggests the following strategy 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 $D(\mathcal{A}||\mathcal{B})$. According to (G.2), the entropy of $\mathcal{A}$ itself can be estimated by measuring the average code length of an optimal code. This is the rationale underlying the methods proposed by Benedetto et al. [1] and Ziv and Merhav [38]. However, to use this idea for general sources (not simply for the memoryless ones that we have considered in this paragraph 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].

G.3.3 Data Compression with the Lempel-Ziv Algorithm

The well-known LZ77 and LZ78 are two seminal universal lossless compression algorithms, introduced by Ziv and Lempel in 1977 and 1978, respectively [36, 37]. We now briefly describe LZ77, which is particularly simple and has become popular as one of the standard algorithm for compression of computer files, due to its speed and efficiency.

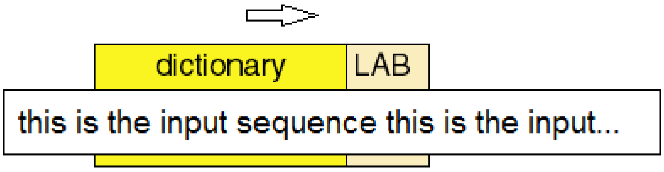


Figure G.1: Sliding window buffer with the dictionary and look ahead buffer (LAB) of the lossless data compressions algorithm LZ77.

The LZ77 algorithm observes the input sequence through a sliding window buffer as shown in Figure G.1. The sliding window buffer consist of a dictionary and a *look ahead buffer* (LAB). The dictionary holds the symbols already analyzed and the LAB the next symbols to be analyzed. At each step, the algorithm tries to express the sequence in the LAB as a sub-sequence 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.

The LZ77 algorithm is widely used for text compression [26], and in recent years it has also been used for text/sequence classification; for example, Pereira Coutinho and Figueiredo [6] used LZ-based dissimilarity measures to achieve state-of-the-art accuracy in the AA task with the Italian Corpus.

G.4 Compression-based Dissimilarity Measures

In recent years, much work has been done concerning the design and development of information-theoretic dissimilarity measures. Most of the proposed approaches were developed in a Kolmogorov complexity context, but an alternative approach can be based on Shannon's relative entropy, as we will show. We use the following notation: let Σ be a finite alphabet; let $\mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{y} = (y_1, y_2, ..., y_m)$ be two sequences with symbols (strings) from Σ; assume that $()$ is the empty sequence and that $|\mathbf{x}|$ denotes the length of $\mathbf{x}$; the subsequence of $\mathbf{x}$ from between positions i and j is denoted as $x[i, j] = (x_i, ..., x_j)$; finally, $\mathbf{x} \circ \mathbf{y} = (x_1, x_2, ..., x_n, y_1, y_2, ..., y_m)$ denotes the concatenation of $\mathbf{x}$ and $\mathbf{y}$..

G.4.1 The Normalized Compression Distance

One of the best known compression-based dissimilarity measure for text is the Normalized Compression Distance (NCD), proposed by Li et al. in 2004 [15]. NCD approximates the non-computable Kolmogorov complexity of a string **x** by the length of a compressed version of **x**, using off-the-shelf compressors, such as *gzip* or *bzip2*, and it is defined for any pair of strings **x** and **y** as

$$NCD(\mathbf{x}, \mathbf{y}) = \frac{C(\mathbf{x} \circ \mathbf{y}) - \min\{C(\mathbf{x}), C(\mathbf{y})\}}{\max\{C(\mathbf{x}), C(\mathbf{y})\}}, \tag{G.5}$$

where $C(\mathbf{x})$ is the length of string **x** after being compressed by a lossless compression algorithm, and $C(\mathbf{x} \circ \mathbf{y})$ represents the length after compression of the concatenation of strings **x** and **y**. The authors demonstrate that it is a metric and claim that it minorizes every computable distance in the class. NCD ranges from 0 to 1 + ϵ, where 0 corresponds to **x** and **y** being identical, and 1 means maximum dissimilarity. The constant ϵ is an upper bound due to imperfections in the compression algorithms, but is unlikely to be above 0.1 for most standard compressors [15]. Although some standard compression algorithms, such as LZ77, LZ78, and even PPM (*prediction by partial matching*), are not guaranteed to satisfy these bounds, NCD has been successfully used in clustering applications [5].

G.4.2 The Ziv-Merhav Relative Entropy Estimate

Ziv and Merhav in 1993 [38] introduced a method for measuring relative entropy between pairs of sequences of symbols, which can be used as a dissimilarity measure for *universal* classification. The method is based on the incremental Lempel-Ziv (LZ) parsing algorithm [37] and on a variation thereof, known as cross-parsing.

The incremental LZ parsing algorithm is a self-parsing procedure of a (length n) sequence **z** into $c(\mathbf{z})$ distinct phrases, such that each phrase is the shortest sequence that is not a previously parsed phrase. For example, with $n = 11$ and $\mathbf{z} = (01111000110)$, the self incremental parsing yields $\{0, 1, 11, 10, 00, 110\}$, thus $c(\mathbf{z}) = 6$. Ziv and Merhav [38] also defined a cross-parsing algorithm that is the sequential parsing of a sequence **z** with respect to another sequence **x**. In this case, $c(\mathbf{z}|\mathbf{x})$ denotes the number of phrases in **z** with respect to **x**. For example, with **z** as above and $\mathbf{x} = (10010100110)$, parsing **z** with respect to **x** yields the set of phrases $\{011, 110, 00110\}$, thus $c(\mathbf{z}|\mathbf{x}) = 3$. Combining these two algorithms, Ziv and Merhav [38] proposed an estimate of the relative entropy between two ergodic sources producing the sequences **z** and **x**, which can be used as a dissimilarity measure between those sequences. Specifically, Ziv and Merhav [38] proved that for two finite order (of any order) Markovian sequences of length n the quantity

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

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 that $(1/n)\, c(\mathbf{z})\, \log_2 c(\mathbf{z})$ is the measure of the complexity of the sequence **z** obtained by self-parsing, thus providing an estimate of its entropy according to (G.2), 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**. The difference between the two quantities does provides a measure of how different the distributions that produced the two sequences are.

G.4.3 The Cross-Parsing Distance

The use of of the Ziv-Merhav relative entropy estimate is not directly applicable in some scenarios, namely because it is defined for sequences of the same length n. When generalizing this definition to sequences of different lengths, several problems arise, with the size of the "model" sequence **x** having a significant impact [10]. To overcome this difficulty, Helmer et al. [10] recently introduced the *cross-parsing distance* (CPD), which is a semi-metric (i.e., of all the conditions that have to be satisfied by a metric, it only does not satisfy the triangle inequality) defined for any pair of strings **x** and **y** (of length respectively $|\mathbf{x}|$ and $|\mathbf{y}|$), as

$$\mathrm{dist}_{CPD}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\left(\frac{|s(\mathbf{x}|\mathbf{y}) \setminus \{\,\mathbf{y}\}\,|}{|\mathbf{x}|} + \frac{|s(\mathbf{y}|\mathbf{x}) \setminus \{\,\mathbf{x}\}\,|}{|\mathbf{y}|}\right), \tag{G.7}$$

where $s(\mathbf{x}|\mathbf{y})$ denotes the multiset of all phrases resulting from the cross-parsing of **x** with respect to **y** and $s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\}$ denotes the removal of a single instance of **y** from the multiset $s(\mathbf{x}|\mathbf{y})$ (if one exists, even if multiple copies exist). If the first not yet parsed symbol in **x** is not found in **y**, then the parsing is simply the symbol itself. For example, if $\mathbf{x} = (ababacbaba)$ and $\mathbf{y} = (aba)$, then $s(\mathbf{x}|\mathbf{y}) = \{aba, ba, c, ba, ba\}$, $s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\} = \{ba, c, ba, ba\}$, and $|s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\}| = 4$. Notice that if $\mathbf{x} = \mathbf{y}$, then $s(\mathbf{x}|\mathbf{y}) = \{\mathbf{y}\}$ and $s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\} = \emptyset$, thus $|s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\}| = 0$; of course, the reciprocal is true, thus $\mathbf{x} = \mathbf{y}$ implies that $\mathrm{dist}_{CPD}(\mathbf{x},\mathbf{y}) = 0$. In contrast, if no symbol in **x** can be found in **y**, then $s(\mathbf{x}|\mathbf{y}) = \{x_1, x_2, ..., x_{|\mathbf{x}|}\}$ (where x_i denotes the i-th symbol of string **x**), thus $s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\} = s(\mathbf{x}|\mathbf{y})$ and $|s(\mathbf{x}|\mathbf{y}) \setminus \{\mathbf{y}\}| = |\mathbf{x}|$; consequently, in this case, $\mathrm{dist}_{CPD}(\mathbf{x},\mathbf{y}) = 1$.

A normalized cross-parsing function defined as $\frac{c(\mathbf{x}|\mathbf{y})}{|\mathbf{x}|}$, where $c(\mathbf{x}|\mathbf{y})$ denote the number of phrases resulting from the cross-parsing of **x** with respect to **y** (as defined in Subsection G.4.2), was successfully used as a dissimilarity measure in an ECG-based biometric recognition task [7]. Thus, we propose to use in this paper as a variant of CPD , a modified version of dist_{CPD} which we call ***CPdist***, defined as

$$\mathrm{CP}_{dist}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\left(\frac{c(\mathbf{x}|\mathbf{y})\; - 1_{\mathbf{x}=\mathbf{y}}}{|\mathbf{x}|} + \frac{c(\mathbf{y}|\mathbf{x})\; - 1_{\mathbf{y}=\mathbf{x}}}{|\mathbf{y}|}\right), \tag{G.8}$$

where $c(\mathbf{x}|\mathbf{y})$ is the number of phrases resulting from the cross-parsing and 1_A is the indicator function of the proposition A.

G.4.4 Incremental Cross-Parsing Algorithm Variant

An implementation of the cross-parsing algorithm based on the LZ77 algorithm (here termed CP77) was proposed by Pereira Coutinho and Figueiredo [6]; this implementation uses a 2 Mbyte dictionary and a 256 byte LAB, where the dictionary is static and only the LAB slides over the input sequence, as shown in Figure G.2. However LZ77 is itself an incremental parsing algorithm; thus, following the same idea, a new implementation is proposed, based on our first implementation but using incremental dictionary updates. In this section, we describe this implementation of the LZ77-based incremental cross-parsing algorithm (termed CP77inc), which we propose for computing text dissimilarity measures.

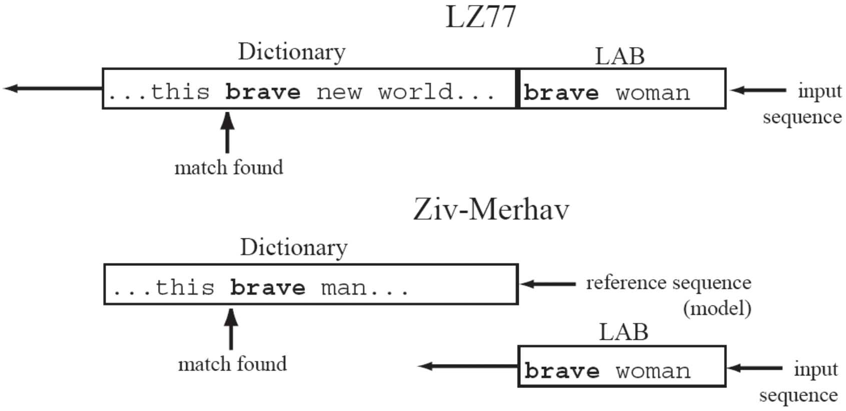


Figure G.2: The original LZ77 sliding window and our modified implementation for cross-parsing, where the dictionary is static and only the look ahead buffer (LAB) slides over the input sequence.

First, let us recall the LZ77 self-parsing procedure of the sequence **x** with length n: initialize a dictionary to the alphabet Σ; assume to have encoded $x[1,i]$; let **s** be the longest prefix of $x[i+1,n]$ that has an occurrence starting at some offset $j \leq i$, with length $|\mathbf{s}|$, and let $x[i+|\mathbf{s}|+1] = a$ be the innovation symbol; then, append to the encoding the triplet $\langle j, |s|, a\rangle$, and repeat the process starting at $x[i+|\mathbf{s}|+2]$. Optimal encoding of the triplet $\langle j, |\mathbf{s}|, a\rangle$ requires $O(log_2(i)+log_2(n)+log_2(|\Sigma|))$ binary digits, and that may result in compression at the end. Furthermore, this procedure implementation takes $O(n)$ steps.

LZSS [31] is a modified version of LZ77, with the following differences: (i) a one-bit flag is used to indicate whether the next encoded prefix is a literal (byte) or a dictionary reference; (ii) if the length of the encoded prefix is less than a “break even” threshold, its symbols are encoded as literals. So, basically, the triplet $\langle j, |s|, a\rangle$ is replaced by the pair $\langle j, |s|\rangle$, and the innovation symbols $\langle a\rangle$ are encoded as literals [19]. According to Ziv and Merhav [38], counting the number of pairs and literals, $c(n)$, is enough for relative entropy estimation, so the encoding process is also discarded from our CP77inc implementation, which itself is based on Mark Nelson’s LZSS implementation [19].

Efficient prefix search is obtained by using a suffix tree [34], which is a data structure that stores all the different suffixes of a string in a way that allows for (fast) substring search in linear time (allows checking if **s** is a substring in $O(|\mathbf{s}|)$ time). Moreover, building a suffix tree for a given string in linear time (i.e., $O(|\mathbf{x}|)$) is also possible using the algorithm proposed by Ukkonen [32]. This is an online algorithm where the suffix tree is constructed on the fly while parsing the string. When moving from symbol x_i to x_{i+1} in a string **x** during parsing, all the suffixes for the string from x_1 to x_i already stored in the tree are extended by x_{i+1}. For our CP77inc implementation, we use the suffix-tree-based sliding window code provided by Larsson [14].

Algorithm 1 illustrates how CP77inc uses a suffix tree with an LZSS-based algorithm to incre-

Algorithm 2 CP77inc: Incremental Cross-Parsing Procedure

Input: z: $1 \times n$ vector, containing the unknown sequence with n symbols.
x: $1 \times m$ vector, containing the model sequence with m symbols.
WINDOWSIZE: an integer constant, setting the sliding window size.
LOOKAHEADSIZE: an integer constant, setting the size for the look ahead buffer (lab).

Output: c_{zx}: an integer, denoting the number of phrases in z with respect to x.

```
initialize suffix tree based sliding window with WINDOWSIZE ;
$lab_z \leftarrow z[1, \mathrm{LOOKAHEADSIZE}]$, $lab_x \leftarrow x[1, \mathrm{LOOKAHEADSIZE}]$ ;
$i \leftarrow 1$ , $j \leftarrow 1$ , $c_{zx} \leftarrow 0$ ;
$Dx \leftarrow (\,)$ ;                                              { // empty dictionary }
while $i \leq |z|$ do
    $c_{zx} \leftarrow c_{zx} + 1$;
    { // cross-parsing of $z$ given $x$ }
    find prefix with largest $len$ so that $lab_z[1, len]$ can be found in $Dx$ ;
    if match not found then
        $len \leftarrow 1$ ;                                        { // literal }
    end if
    $lab_z \leftarrow \mathrm{updateLAB}(z, i, len)$;
    $i \leftarrow i + len$ ;
    { // dictionary update using self parsing over $x$ }
    if $j \leq |x|$ then
        find prefix with largest $len$ so that $lab_x[1, len]$ can be found in $Dx$ ;
        if match not found then
            $len \leftarrow 1$ ;
        end if
        $Dx \leftarrow Dx \circ x[j, j + len - 1]$ ;                { // append match to dictionary }
        add to suffix tree the prefix found $x[j, j + len - 1]$ ;
        $lab_x \leftarrow \mathrm{updateLAB}(x, j, len)$ ;
        $j \leftarrow j + len$ ;
    end if
end while
return $c_{zx}$ ;
```

mentally cross-parse strings in linear time. The cross-parsing of string $\mathbf{z}$ with respect to the string $\mathbf{x}$ involves some details, which we now briefly describe; it uses one sliding window to hold both the dictionary D_x and the LAB lab_x of the model string $\mathbf{x}$. In addition, it uses another sliding window (smaller) to hold the LAB lab_z for the unknown string $\mathbf{z}$. Notice that the dictionary D_x is empty at the beginning. Then, a two-step loop is repeated until the end of $\mathbf{z}$ is reached: the cross-parsing of $\mathbf{z}$ given $\mathbf{x}$; the self parsing of $\mathbf{x}$ including dictionary update as long as $\mathbf{x}$ lasts. This makes sequences of different lengths allowed, by stopping the dictionary update whenever the end of $\mathbf{x}$ is reached and keep using it as a "static" dictionary. Every time the loop is executed a counter c_{zx} is incremented. Finally, we call the method of relative entropy estimate via definition (G.6) as ***ZMM*** when based on CP77, while the method that uses the proposed algorithm CP77inc we call ***ZMMinc***.

G.5 Dissimilarity-Based Classification

At the core of dissimilarity-based methods for classification is the computation of pairwise dissimilarities between the object (*e.g.* text) to be classified and a set of (or all) objects (*e.g.*, texts) in the training set. Of course, there are several ways to use dissimilarity values to define a classifier, the simplest of which is arguably to use a k-NN classifer; in this case, the object to be classified is simply assigned to the majority class in its k nearest (in the adopted similarity measure) neighbors (with some rule to break ties). A more sophisticated approach is offered by the dissimilarity space approach [22, 23], which uses the dissimilarity values as features that characterize the object to be classified, based on which several different types of classifiers can be used, namely k-NN in the dissimilarity space or support vector machines (SVM).

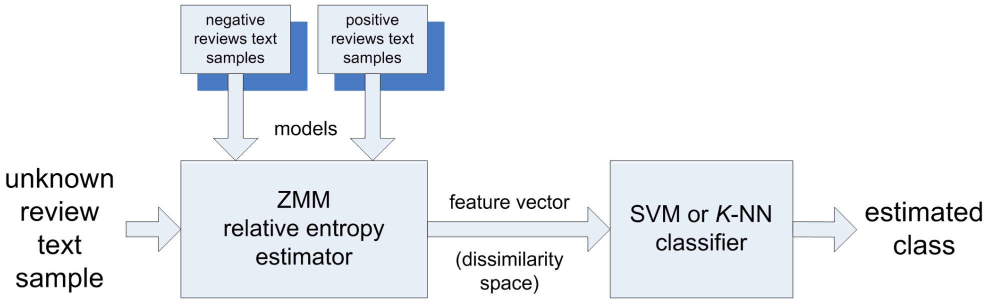

negative reviews text samples
positive reviews text samples
models
unknown review text sample
ZMM relative entropy estimator
feature vector
(dissimilarity space)
SVM or K-NN classifier
estimated class

Figure G.3: Block diagram of the proposed system for sentiment analysis.

In this paper, we propose to use a dissimilarity space approach, where the representations are built by using the dissimilarity/distance measures described in the previous section (see Figure G.3). Once in possession of a dissimilarity-based representation of a training set, any standard classification method can be used; in this paper, we report results based on k-NN and (linear) SVM classifiers.

Formally, let us consider a training set of objects (texts) $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, where each object belongs to some set $\mathcal{X}$ (*e.g.*, the set of finite length strings of some finite alphabet Σ), and some dissimilarity measure between pairs of objects, $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In the dissimilarity-based approach, each object (either in the training set or a new object to be classified after training) is represented by the vector of its dissimilarities with respect to the elements of $\mathbf{X}$ (or a subset thereof). That is, the training set in the so-called dissimilarity space becomes

$$\mathcal{D} = \{\mathbf{d}_1, ..., \mathbf{d}_n\},$$

where

$$\mathbf{d}_i = \begin{bmatrix} D(\mathbf{x}_i, \mathbf{x}_1) \\ \vdots \\ D(\mathbf{x}_i, \mathbf{x}_n) \end{bmatrix} \in \mathbb{R}^n.$$

An important aspect of dissimilarity-based approaches is that very few conditions are put of the dissimilarity measure; namely, it doesn't have to be a metric, it doesn't even need to be symmetric [22, 23]. Dissimilarity representations can also be based on a subset of the training set (rather than all of it), in which case the dissimilarity space has dimension equal to the cardinal of that subset; some work has been devoted to methods for selecting this subset [24].

G.6 Experiments

We assessing the accuracy of the proposed methods on the two different text classification problems mentioned above: *sentiment analysis* (SA), which is a binary problem, and *authorship attribution* (AA), which is a multi-class problem. In this assessment, we use several corpora.

G.6.1 Experimental Datasets and Setup

We conduct our SA experiments on five well known datasets. Namely, we use the Movie Review Data[6] (more precisely the polarity dataset v2.0), introduced by Pang and Lee [20], and the Multi-Domain Sentiment Dataset (version 2.0)[7], introduced by Blitzer et al. [2], which includes four datasets with Amazon reviews of four classes of products: Books, DVD, Electronics, and Kitchen. Each of the five datasets are labeled by humans and include 1,000 positive and 1,000 negative unprocessed reviews. We report 5-fold cross-validation (CV) accuracy estimates, following the same protocol of Xia et al. [35], where in each run, 1600 examples are used to train and 400 examples to test.

For the AA experiments, we use three different corpora, briefly described in section G.2: (i) the Italian Corpus, introduced by Benedetto et al. [1], with 90 texts from 11 Italian authors[8] spanning the 13th to 20th century; (ii) the English Corpus, introduced by Ebrahimpour et al. [9], containing 168 short stories by seven undisputed English authors[9] from the late 19th century to the early 20th

[6] Available at `http://www.cs.cornell.edu/people/pabo/movie-review-data`

[7] Available at `http://www.cs.jhu.edu/~mdredze/datasets/sentiment`

[8] Available at `http://www.liberliber.it`

[9] Available at `http://promo.net/pg`

century, truncate to approximately the first 5,000 words, due to the differing lengths of the books; (iii) the Federalist Papers [17], where we consider only the 70 undisputed (out of 85) political essays published in 1788 by three American authors[10]. Due to the reduced number of texts samples in each corpus, we use leave-one-out cross-validation (LOO-CV) to assess the accuracy of the classifiers. Table G.1 shows some statistics about the used corpora.

Table G.1: Some (statistical) facts about the three corpora used for AA performance evaluation. Notice that **sa** denotes *samples*, **cl** denotes *class*, **mLen** denotes *mean length* and **kB** denotes *kByte*.

Corpus	**No. sa**	**Max sa/cl**	**Min sa/cl**	**mLen**	**Max mLen/cl**	**Min mLen/cl**
Italian	90	15	4	343 kB	727 kB	158 kB
English	168	26	14	38 kB	44 kB	34 kB
Fed. Papers	70	51	5	13 kB	17 kB	10 kB

We stress that, in all the experiments, we do not use any text preprocessing. The k-NN and SVM classifiers (with linear kernel) used are implemented by the PRTools Matlab toolbox for pattern recognition [11] (version 4). The SVM penalty parameter (usually denoted by C) value was set to 1 or adjusted by CV. Reported results are in terms of the classification accuracy, expressed in percentage.

G.6.2 Experimental Results

Table G.2 shows the 5-fold CV accuracy estimates of an SVM (with $C = 1$, except in the case denoted as ZMMoptC) working in the dissimilarity space, using as dissimilarity measures the above described NCD, CPdist and ZMM. For comparison purposes, we also show the baseline and best results on the same datasets, described by Xia et al. [35]. Our method ZMMoptC achieves an accuracy of 82.41%, which is better than both baselines and is close to the best results reported by Xia et al. [35].

Regarding AA, our experiments were done with both k-NN and SVM classifiers on the dissimilarity space, with the accuracy assessed by LOO-CV. Table G.3 shows the accuracy results for each of the corpus when using NCD, CPdist and ZMMinc as dissimilarity measures. Our method ZMMinc, with optimized C, obtains an accuracy of 98.8% on the English Corpus (only fails 2 out 168 texts), outperforming the methods of Ebrahimpour et al. [9]; on the two other corpora, the performance is approximately 4% below the baseline. Notice, however, that our results are obtained without any feature design/engineering or any text preprocessing, thus can be considered as highly competitive with those other methods.

[10] Available at `https://github.com/matthewberryman/author-detection/tree/master/Federalist%20Texts`

[11] Available at `http://www.prtools.org/index.html`

Table G.2: SA results: 5-fold CV accuracy percentages, using several dissimilarity measures with SVM classifiers on 5 benchmark datasets. For comparison, we also show in the last four columns the results obtained by Xia et al. [35] over the same datasets, using the approaches POS-based (M1), part-of-speech information, and WR-based (M2), word relation features, plus the baselines assumed by those authors, respectively.

Dataset	**NCD**	**CPdist**	**ZMM**	**ZMMoptC**	**Baseline1**	**M1**	**Baseline2**	**M2**
Movies	84.85	80.45	84.60	85.80	84.75	86.80	86.45	87.70
Books	74.65	79.15	78.85	80.85	74.70	80.10	77.65	81.80
DVD	78.05	79.60	79.05	81.95	77.20	80.40	79.45	83.80
Elec	81.60	80.85	78.05	81.25	80.05	83.40	82.50	85.95
Kitchen	82.10	81.40	78.70	82.20	83.25	84.90	85.40	88.65
Average	**80.25**	**80.29**	**79.85**	**82.41**	**79.99**	**83.12**	**82.29**	**85.58**

Table G.3: AA results: leave-one-out cross-validation accuracy percentages, using several dissimilarity measures with SVM classifiers on 3 benchmark coprpura.

		K-NN			**SVM**		
Corpus	**Baseline**	**NCD**	**CPdist**	**ZMMinc**	**NCD**	**CPdist**	**ZMMinc**
Italian	97.8	52.2	82.2	64.4	80.0	92.2	**94.4**
English	96.4	84.5	91.7	87.5	95.2	95.2	**98.8**
Federalist	97.1	82.9	90.0	84.3	62.2	81.4	**92.9**

G.7 Conclusions

A central and crucial task in text classification is to choose an appropriate set of features. Achieving good accuracy usually requires careful feature engineering and complex preprocessing stages, which may become prohibitive in the emerging context of classification massive sets of electronic texts, available from Internet and others sources. In this paper, we proposed methods for automatic text classification using information-theoretic dissimilarity measures, based on universal data compression algorithms, which bypass the feature design and preprocessing stages. The proposed methods map the raw texts into a feature space vectors, using *universal* dissimilarity measures.

Experiments were done using several dissimilarity measures for evaluating the proposed methods on two classical text classification problems: sentiment (polarity) analysis, which is a binary problem, and authorship attribution, which is a multi-class problem. We tested k-NN and SVM classifiers, with the best results achieved with the latter. Experimental results reveal that the proposed methods approximate, or even outperform in some cases, previous state-of-the-art techniques, despite being much simpler, in the sense that they do not require preprocessing and feature engineering. In future work, we will aim at obtaining even better results, by using other kernels, other dissimilarity representations, and by exploiting the possibility of selecting a subset of objects with respect to which the dissimilarity representations are obtained [24].

References

[1] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. Language Trees and Zipping. *Physical Review Letters*, 88(4):048702, January 2002.

[2] John Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, 2007.

[3] Daniele Cerra and Mihai Datcu. Algorithmic Cross-Complexity and Relative Complexity. *2009 Data Compression Conference*, pages 342–351, March 2009.

[4] Gregory J. Chaitin. On the Length of Programs for Computing Finite Binary Sequences: Statistical Considerations. *Journal of the ACM*, 13:547–569, 1969.

[5] Rudi Cilibrasi and P. M. B. Vitányi. Clustering by compression. *Information Theory, IEEE Transactions on*, 51(4), 2005.

[6] David Pereira Coutinho and Mário A. T. Figueiredo. Information Theoretic Text Classification Using the Ziv-Merhav Method. In *Pattern Recognition and Image Analysis. Springer Berlin Heidelberg*, volume 1, pages 355–362, 2005.

[7] David Pereira Coutinho, Ana L. N. Fred, and Mário A. T. Figueiredo. Personal Identification and Authentication based on One-lead ECG using Ziv-Merhav Cross Parsing. In *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems (PRIS 2010)*, pages 15–24, 2010.

[8] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[9] Maryam Ebrahimpour, Tlis J. Putniš, Matthew J. Berryman, Andrew Allison, Brian W.-H. Ng, and Derek Abbott. Automated authorship attribution using advanced signal classification techniques. *PloS one*, 8(2):e54998, January 2013.

[10] Sven Helmer, Nikolaus Augsten, and Michael Böhlen. Measuring structural similarity of semistructured data based on information-theoretic approaches. *The VLDB Journal*, 21(5):677–702, February 2012.

[11] A. N. Jebaseeli and E. Kirubakaran. A Survey on Sentiment Analysis of(Product) Reviews. *International Journal of Computer Applications*, 47(11):36–39, 2012.

[12] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, 1998.

[13] A. N. Kolmogorov. Three approaches to the quantitative definition ofinformation'. *Problems of information transmission*, 1(1):3–11, 1965.

[14] N. J. Larsson. *Structures of string matching and data compression.* Phd thesis, Lund University, Sweden, 1999.

[15] M. Li, Xin Chen, and Xin Li. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.

[16] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer Verlag, 1997.

[17] F. Mosteller and D. Wallace. *Inference and disputed authorship: The Federalist.* Addison-Wesley, 1964.

[18] F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[19] Mark Nelson and Jean-loup Gailly. *The Data Compression Book.* M&T Books, New York, 2nd editio edition, 1995.

[20] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics*, page 271, 2004.

[21] Bo Pang, Lillian Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86, 2002.

[22] E. Pekalska, P. Paclik, and R. P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.

[23] Elbieta Pkalska and Robert P. W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, 23(8):943–956, June 2002.

[24] Elbieta Pkalska, Robert P. W. Duin, and Pavel Paclík. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, February 2006.

[25] A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani. Data compression and learning in time sequences analysis. *Physica D: Nonlinear Phenomena*, 180(1-2):92–107, June 2003.

[26] David Salomon and Giovanni Motta. *Handbook of Data Compression.* Springer, 2010.

[27] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[28] R. J. Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1, March 1964):1–22, 1964.

[29] R. J. Solomonoff. A formal theory of inductive inference. Part II. *Information and control*, 7(2, June 1964):224–254, 1964.

[30] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[31] J. A. Storer and T. G. Szymanski. Data compression via textual substitution. *Journal of the ACM (JACM)*, 29(4):928–951, 1982.

[32] E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260, September 1995.

[33] G. Vinodhini and R. M. Chandrasekaran. Sentiment Analysis and Opinion Mining: A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6), 2012.

[34] Peter Weiner. Linear pattern matching algorithms. In *14th Annual Symposium on Found. of Computer Science (FOCS), Iowa City, Iowa*, pages 1–11, 1973.

[35] Rui Xia, Chengqing Zong, and Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6):1138–1152, March 2011.

[36] J. Ziv and a. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977.

[37] J. Ziv and a. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, September 1978.

[38] J Ziv and N Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39(4):1270–1279, 1993.