Multiresolution Similarity Search in Time Series Data: An Application to EEG Signals

Amalia Charisi[1], Fragkiskos D. Malliaros[2], Evangelia I. Zacharaki[1], Vasileios Megalooikonomou[1,3]

[1]Department of Computer Engineering and Informatics, University of Patras, Greece
[2] Computer Science Laboratory, École Polytechnique, France
[3] Center for Data Analytics and Biomedical Informatics, Temple University, PA, USA

{charisa, vasilis}@ceid.upatras.gr, fmalliaros@lix.polytechnique.fr, ezachar@upatras.gr

ABSTRACT

Time series constitute a prevalent data type that arise in several diverse disciplines (e.g., biomedical data, sensor data, images, video data), and therefore analyzing time series is a significant task with a plethora of important applications. In this paper, we study the general problem of similarity search in time series databases and we propose a novel multiresolution indexing (i.e., representation) and retrieval method for time series similarity search. Our approach is motivated by the idea that if we examine a time series at different resolution levels, we could possibly acquire further insights about the data. The proposed algorithm adopts a combined, two-step pruning (filtering) strategy to further reduce data dimensionality by discarding irrelevant time series (i.e., false alarms). At a first level, the time series are represented by line segments and filtered by the triangular inequality property. Then, a Vector Quantization like scheme is applied to encode data and thus to reduce dimensionality.

We test and demonstrate the performance of the proposed method, analyzing EEG time series data for retrieval of one of the constituent brain waveforms in EEG recordings, the K-complex, but the method can as well be applied for retrieval of other patterns of interest in time series analysis. The automatic detection and categorization of the EEG patterns will allow the advanced correlation analysis of large amounts of data and will lead to advanced decision making capabilities assisting diagnosis by medical professionals.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data mining*; H.2.4 [**Information Systems**]: Systems—*Multimedia databases*

PETRA '13, May 29-31, 2013, Island of Rhodes, Greece.

General Terms

Algorithms, Experimentation

Keywords

Time series, Similarity search, EEG signals, Assistive Environments

1. INTRODUCTION

A time series is a sequence of real numbers taken by the observation of a variable over time. Time series can be found in several domains including medical, financial and business. For this reason, time series analysis and its applications have received a lot of attention over the last years.

An important application of time series data analysis involves brain and body activity monitoring using a variety of modalities, such as EEG, ECG, EOG, EMG. EEG is the major modality that is used in the literature for several purposes including diagnosis and seizure detection in epilepsy studies. Specifically in such applications, it is also crucial to analyze time series of sleep EEG due to the fact that several sleep factors, mechanisms, or conditions were found to aggravate seizures [12].

Here, we deal with the problem of similarity search in time series databases. A similarity search query (range query) is defined as follows: given a time series database T and a query (q, r), where r represents a threshold, find all the time series $t \in T$ that their distance from q is smaller than r. The easiest but the most expensive way to answer such a query is to scan the database sequentially. To improve the performance of similarity searching, several dimensionality reduction techniques have been suggested in the literature; these techniques, transform time series into a lower dimensional space and then process more efficiently the similarity search task in this reduced space.

In this paper, we propose MR-PVQ (MultiResolution Piecewise Vector Quantization), a representation and similarity search method for time series datasets. In general, it would be helpful if we are able to examine a time series at different resolution levels, in order to acquire further insights about the data. Our method uses line segments to represent time series and then a dimensionality reduction technique based on a vector quantization method. The above scheme is applied in multiple resolution levels, discarding the non-qualifying sequences in every level.