Neurocomputing 129 (2014) 556–569

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# A new similarity measure based on shape information for invariant with multiple distortions

Xiaoxu He [a,*], Chenxi Shao [a,b], Yan Xiong [a]

[a] School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China
[b] Anhui Province Key Laboratory of Software in Computing and Communication, Hefei 230027, China

ARTICLE INFO

*Article history:*
Received 2 September 2012
Received in revised form
10 September 2013
Accepted 23 September 2013
Communicated by P. Zhang
Available online 18 October 2013

*Keywords:*
Time series
Similarity search
Distortion
Data mining

ABSTRACT

Due to the characteristics of noise and volatility, two similar time series always appear in diverse kinds of distortions, which usually are considered as the combinations of the following basic transformations: noise, amplitude shift, amplitude scaling, temporal scaling, and linear drift. In this paper, a novel similarity measure (*SIMshape*) invariant to these basic distortions and any combinations of them is proposed. It is parameter-free and easy to implement. Specifically, a multi-scale shape approximation for time series based on Discrete Haar Wavelet Transform, key point extraction and symbolization is presented first; then, based on this proposed representation and a scale-weight factor, a robust similarity measure is proposed. The novelty of *SIMshape* lies in two aspects as follows: (a) symbolizing key points sequence extracted from approximate wavelet coefficients; (b) adding the scale-weight factor and shape similarity in the similarity criterion. To show the effectiveness and efficiency, *SIMshape* is compared with other popular methods Euclidean Distance (ED), LB_keogh, Complexity Invariant Distance (CID), and ASEAL (Approximate Shape Exchange ALgorithm) using two indices: the number of kinds of distortions and the degree of distortion. Obtained results show that compared with ED, CID, LB_keogh, and ASEAL, *SIMshape* has better robustness in synthetic data, and shows better performance in real time series classification.



## 1. Introduction

The research on similarity measure is one of the core aspects in time series data mining [1–3]. Almost every task of time series data mining requires a subtle notion of similarity between series [1–3,5]. Due to the characteristics of noise and volatility, the two similar time series always appear in diverse kinds of distortions, which are usually seen as the combinations of the following five basic distortions: noise, amplitude scale, amplitude shifting, temporal scaling, and linear drift [3,4,6–8].

In recent years, hundreds of techniques have been designed to study the similarity measure with invariance under the mentioned basic distortions for time series [3,4]. As a result, the similarity model has been extended in many different directions [3]: taking time warping into account [7–9,11,16–22], allowing amplitude shifting [7,8,22], allowing time series of different lengths [7–9,11,16–22], tolerating some degree of noise [7,8,10,11,15–18,21], and invariant to the complexity [6].

Many literatures often use the number of kinds of tolerable distortions to measure the performance of similarity measure [3,6,8]. The more kinds of distortions a similarity model tolerates, the more powerful the similarity model is. For example, according to the study results of [7,8], ASEAL has been demonstrated to be superior to others used in the literature on ECG datasets for dealing with four basic distortions: noise, offset translation, amplitude scaling and time axis scaling. However, there are few literatures discussing the tolerable degree of a specific distortion in the evaluation of similarity measure; while in this paper, it is considered as an important index to measure the performance of similarity measure. Besides, most of the existing approaches take a toll to tune its parameters and compute [10,11,15,16,18,20–22]. For instance, for EDR and LCSS measures [15,16,18], a threshold parameter is required to be set, which is difficult without a priori knowledge of the data.

Inspired by shape recognition, a novel similarity measure *SIMshape* is introduced to address multiple distortions in this paper. The number of kinds of distortions and the degree of distortion are used to measure the robustness of *SIMshape*. In order to provide a comprehensive validation, the experiments on the synthetic data and five real time series data from different domains have been conducted. The major contributions of this paper are the following:

- To record the most salient features of the original time series from different scales, a new symbolic approximate representation for time series is introduced. The representation is obtained through Multi-scale Discrete Haar Wavelet Transform, key point extraction, and symbolization. Unlike the traditional approximation methods,

* Corresponding author. Tel.: +86 551 63606694; fax: +86 551 3601583.
*E-mail address:* xiaoxuhe@mail.ustc.edu.cn (X. He).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.09.003

such as Discrete Fourier Transform (DFT) [23], Symbolic Aggregate Approximation (SAX) [24,25] and Piecewise Aggregate Approximation (PAA) [27], it does not need to preset any parameter. The symbolization technique significantly reduces dimensionality. The essential characteristics of the original time series is retained by the application of Multi-scale Discrete Haar Wavelet Transform and key point extraction retain. Therefore, the multi-scale shape information assures the efficiency and effectiveness of *SIMshape*.

- To improve *SIMshape* robustness to various transformations, the scale-weight function for *SIMshape* is designed. It makes *SIMshape* emphasize the basic shape information of the original sequence, which is preserved in the coarse level. As the essential characteristics of time series are not altered by the degree of distortions mentioned, so the distortions have relatively little impact on the information in the coarse level. As a result, the robustness of *SIMshape* is improved by assigning bigger weighted values to the coarse level.
- To measure the similarity between two time series, a novel similarity measure *SIMshape*, based on the multi-scale shape information and the scale-weight function, is presented. *SIMshape* is parameter-free and easy to implement. Through a set of objective tests on synthetic data sets and five real time series data sets from different application domains, it can be found that *SIMshape* is more robust to various deformations than LB_keogh [10,11], ED, CID [6] and ASEAL [7,8], and more accurate than other four methods when applied to classify real time series.

The rest of this paper is organized as follows. In Section 2, the current known basic distortions for time series are reviewed, and related methods are cited and commented. In Section 3, a new similarity measure and its specific implementation is put forward. In Section 4, experiments in synthetic time series and real-world time series have been conducted to evaluate the proposed method. In Section 5, conclusions and some potential future work are given.

## 2. Related work

### 2.1. Problem statement

Suppose that there are two time series $Q$ and its small distortion $Q_d$, which means that the degree of distortion cannot alter the nature of $Q$. Specifically, $Q$ and $Q_d$ have the same basic shape information ignoring the subtle difference, and they seem very similar to the human eye.

In this paper, the following five basic transformations [3,4,6–8] are considered. As shown in Fig. 1, these basic transformations are defined as follows:

- *Noise*: The noise distortion means that two time series have the same basic shape while one is rough with burrs due to noise.
- *Amplitude shift*: The amplitude shift distortion means two time series with the same basic shape fluctuate in different levels of baseline.
- *Amplitude scaling*: The amplitude scaling distortion denotes two time series with the same basic shape fluctuate in different amplitude. Based on the scaling factor, the amplitude scaling can be divided into two types: if the scaling factor is larger than 1, it is amplitude stretch; if the scaling factor is smaller than 1, it is amplitude shrink.
- *Time scaling*: The time scaling distortion signifies that two time series have the same basic shape, while, for one of them, the width of the waveform scales according to the same proportion. Like the amplitude scaling, based on the scaling factor, time scaling can be divided into two types: if the scaling factor is smaller than 1, it is amplitude stretch; if the scaling factor is larger than 1, it is amplitude shrink.
- *Linear drift*: The linear drift distortion intends that for one of the two time series with the same basic shape, its value is linearly increasing or decreasing under the influence of a certain linear factor.

Due to the disturbance of many factors, the distortions of real time series can be considered as the combination of these five basic transformations [33]. The purpose of introducing these transformations is not actually to perform them, but instead to extend the semantics of similarity to 'ignore' them. In this paper, it is required that the degree of distortions never alters the essential characteristics of the original time series, otherwise it is meaningless to discuss the robust similarity measure. An ideal similarity measure should be invariant to the distortions [3,6]. The more kinds a similarity model tolerates, the more powerful the similarity model is [3].

### 2.2. Existing methods

The ubiquitous Euclidean Distance is regarded as one of the most popular similarity methods when comparing $Q$ and $Q_d$ [1,4].

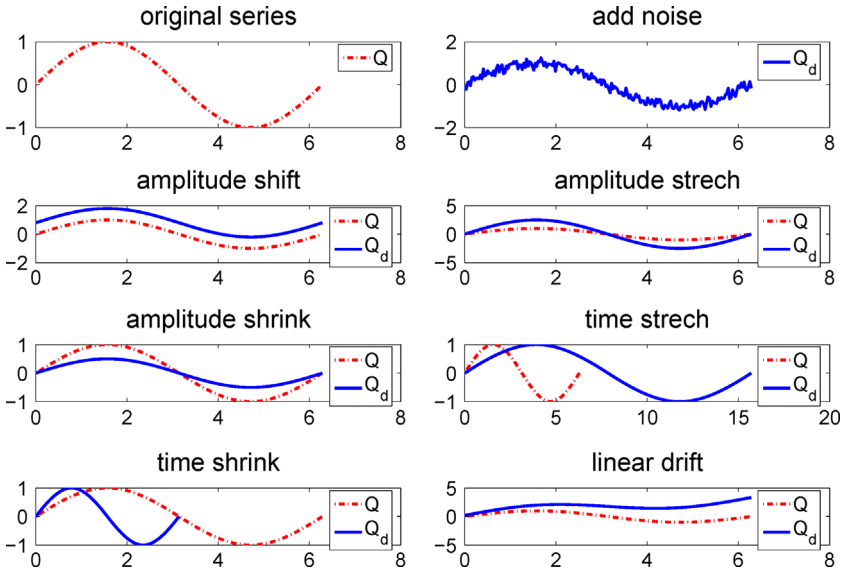


**Fig. 1.** The basic distortions in time series similarity.