International Journal of Engineering

Journal Homepage: www.ije.ir

An Empirical Comparison of Distance Measures for Multivariate Time Series Clustering

A. Salarpour, H. Khotanlou*

RIV Lab, Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran

PAPER INFO

Paper history:
Received 15 October 2017
Received in revised form 06 December 2017
Accepted 21 December 2017

Keywords:
Multivariate Time Series
Similarity Measures
Clustering
Evaluation

ABSTRACT

Multivariate time series (MTS) data are ubiquitous in science and daily life, and how to measure their similarity is a core part of MTS analyzing process. Many of the research efforts in this context have focused on proposing novel similarity measures for the underlying data. However, with the countless techniques to estimate similarity between MTS, this field suffers from lack of comparative studies using quantitative and large scale evaluations. In order to provide a comprehensive validation, an extensive evaluation of similarity measures for MTS data clustering is conducted. Effectivness of fourteen well-known similarity measures and their variants on 23 MTS datasets, coming from a wide variety of application domains, were evaluated experimentaly. In this paper, an overview of these different techniques is given and the empirical comparison regarding their effectiveness based on agglomerative clustering task is presented. Furthermore, the statistical significance tests were used to derive meaningful conclusions. It has been found that all similarity measures are equivalent, in terms of clustering F-measure, and there is no significant difference between similarity measures based on our datasets. The results provide a comparative background between similarity measures to find the most proper method in terms of performance and computation time in this field.

***doi**: 10.5829/ije.2018.31.02b.08*

1. INTRODUCTION

In the last few years, multivariate time series (MTS) data have been appeared extensively in scientific domains [1, 2] that represent valuable information subject to analysis, clustering, classification, indexing, and interpretation [3-5]. Real-world applications include daily fluctuations of the stock market (financial data analysis[6]), electrocardiogram data mining (medical data processing [7]) and moving object identification (motion data analysis [8]). Even object shapes and handwriting data could be transformed to time series data for further analyzing. In addition, multivariate time series datasets are always embedded with additional information such as class labels, place and time of occurrence [9].

A key concept toward dealing with multivariate time series data is determining their pairwise similarity. In fact, an multivariate time series similarity (or dissimilarity) measure is a core routine to many data mining [10], retrieval, clustering, and classification tasks [4, 5, 8]. Furthermore, deriving a distance, that correctly captures semantics and reflects underlying similarity of multivariate time series data, is not straightforward. Apart from challenges related to the high dimensionality of such data, calculation of similarity measure requires to be fast and efficient.

The generalized framework for the task of time series mining encompasses: **data preparation phase** which includes sensing that explains the idea of time series data collection from different sources like human, ECG and stock data. **Pre-processing** step cleans the gathered data from missing values. **Primary data representation** referres to the methods that are used for representing stored information. **Time series analysis** is the most important part of the framework that includes similarity measures and analysis techniques. Similarity measure has the responsibility of calculating the similarity between time series data that plays an essential role for further analysis. Analysis section could include many techniques that categorize the time series

*Corresponding Author's Email: *khotanlou@basu.ac.ir* (H. Khotanlou)

data in an automatic or semi-automatic way, here only two of the more common time series analysis techniques are listed, namely classification and clustering. The last part of the framework, but certainly not the least, is **knowledge discovery** which is totally dependent on the intended application and extracts applicable knowledge from the result of time series analysis phase for further investigation, such as discovery of relationship, characterization of data, time series-based prediction, traffic modelling, and even detecting abnormal activities.

As a result, time series mining has been receiving much attention in the past decade, which resulted in a large number of studies, introducing approaches for querying, classifying, and clustering of time series. However, mining of time series data can be challenging due to the fact that single continuous data may result in countless different discrete time series representations. One of the key aspects for achieving effectiveness and efficiency when analyzing a time series is measuring the similarity of two time series. The similarity measure is a real-valued function which reflects the similarity between time series that could be the inverse of distance function over time series. By contrast, unlike the conventional straightforward distance definition, the definition of distance between time series should precisely quantify the similarity (dissimilarity) between time series which is desirable for retrieval, classification, clustering and other analyzing routines of time series.

In the recent decade, researches on time series similarity measures have become popular. Many techniques have been proposed to measure the time series data similarity. Although literature covers a wide variety of such similarity measures, this work will focus on most cited techniques which emerge repeatedly throughout related works. For example, Euclidean distance (ED) [11], Dynamic Time Warping (DTW) [12, 13], Weighted DTW (WDTW) [2], Longest Common Subsequence (LCSS) [14], Edit Distance on Real sequences (EDR) [15], Edit Distance with Real Penalty (ERP) [16], Sequence Weighted Alignment model (SWALE) [17], Time Warp Edit Distance (TWED) [18], the Move-Split-Merge distance (MSM) [19], Hausdorff, Fréchet [20, 21], Symmetrized Segment-Path Distance (SSPD) [22], derivative DTW (DDTW) [23], and Complexity Invariant DTW (CIDTW) [24] are studied in this work.

Few researchers have addressed the problem of finding the best similarity measure for time series analysis. Preliminary work was carried out by Ding et al. [25], who compared and discussed nine different similarity measures over 38 diverse fixed-length univariate time series datasets for the classification task. There exist some other extensive experimental works in time-series classification that examine a number of similarity measures for the time-series based on 1-nearest neighbor classification task to compare their strength and weakness [26-29]. A key limitation of these works is that all of the comparative studies consider only univariate time series data. Another issue is that all of the cited researchers use classification task for comparing similarity measures.

Unfortunately, even despite some works in the area, it is still unclear which similarity measure is more appropriate for the multivariate time series clustering task. However, with the multitude of competitive techniques, we believe that there is a strong need for a comprehensive comparison of similarity measures in multivariate time series data clustering context that has drawn the most attention from data mining researchers. Every newly proposed similarity measure has claimed a kind of superiority over some of the existing methods. On the other hand, their empirical evaluations have not been the same and perhaps adequate. This has not only confused newcomers and specialists, but also led to the use of a wrong methods based on incomplete and not generalized results.

To address these problems, an empirical evaluation of similarity measures for multivariate time series clustering was performed. As for the considered measures, we decided to include nine elastic similarity measures, as these were found the state-of-the-art similarity measures. Apart from these nine, we chose three geometry-based and two differential-based similarity measures that were not considered in earlier studies. The main contributions of this work can be summarized as follows: 1) an extensive summary and background of the considered similarity measures is presnted with basic formulations. 2) The time efficieny of 14 investigated similarity measures are compared over 23 highly diverse multivariate time series datasets. 3) Similarity measures effectiveness and efficiency are evaluated using agglomorative clustering technique. 4) Statistical significance tests are used to evaluate the superiority of given similarity measures.

2. PRELIMINARIES

Typically, a multivariate time series data is a temporal sequence which is sampled from a continuous signal. For simplicity and without any loss of generality, the multivariate time series data are considered discrete hereafter.

Definition 1. Let X be a set of multivariate instances along the time which forms a multivariate time series data. The multivariate time series dataset is defined as follow:

$$\mathrm{X} = \left\{ X_1, \ldots, X_{n_\mathrm{X}} \right\} \quad (1)$$

where X_k is a multivariate time series instance and n_X

is the number of multivariate time series data in X dataset.

Definition 2. Each multivariate time series data is formally defined as a sequence of pairs:

$$X_k = \left[\left(x_{k,1}, t_{k,1}\right), \ldots, \left(x_{k,i}, t_{k,i}\right), \ldots, \left(x_{k,n_k}, t_{k,n_k}\right)\right], \quad \forall i \in \left[1, \ldots, n_k\right] \tag{2}$$

where each $x_{k,i} \in \mathrm{R}^d$ is an instance in d-dimensional space, each $t_{k,i} \in \mathrm{R}$ is a temporal index at which the corresponding $x_{k,i}$ occurs and n_k is the number of samples in multivariate time series data X_k.

Definition 3. A piecewise linear multivariate time series is a set of line segments that are bounded between successive multivariate time series instances which is defined as:

$$Xs_k = \left(xs_{k,1}, \ldots, xs_{k,i}, \ldots, xs_{k,n_k-1}\right) \tag{3}$$

where each $xs_{k,i} \in \mathrm{R}^{2\times d}$ is a line-segment $\overline{s_{k,i}s_{k,i+1}}$ of Xs_k that is bounded between $xs_{k,i}$ and $xs_{k,i+1}$.

As a basic measure to find the distance between multivariate time series data samples, the Euclidean metric [11] is used in the rest of this paper as follows:

$$d_{eucl}\left(x_{k,i}, x_{l,j}\right) = \sqrt{\sum_{dim=1}^{d} \left(x_{k,i}^{dim} - x_{l,j}^{dim}\right)^2} \tag{4}$$

where $d_{eucl}\left(x_{k,i}, x_{l,j}\right)$ is the Euclidean distance between two d-dimensional time series samples sample $x_{k,i}$ and $x_{l,j}$.

Furthermore, the point-to-segment distance is defined as a minimum Euclidean distance between sample points and given multivariate time series segment as revealed by:

$$d_{p2s}\left(x_{k,i}, xs_{l,j}\right) = \begin{cases} d_{eucl}\left(x_{k,i}, x_{k,i}^{proj}\right) & \textbf{if } x_{l,j}^{proj} \in xs_{l,j} \\ \min\begin{cases} d_{eucl}\left(x_{k,i}, x_{l,j}\right) \\ d_{eucl}\left(x_{k,i}, x_{l,j+1}\right) \end{cases} & \textbf{otherwise} \end{cases} \tag{5}$$

where $x_{k,i}^{proj}$ is the orthogonal projection of multivariate time series data sample $x_{k,i}$ on the multivariate time series segment $xs_{l,j}$ and $d_{p2s}\left(x_{k,i}, xs_{l,j}\right)$ is the point-to-segment distance between $x_{k,i}$ and $xs_{l,j}$.

A similarity measure is a numerical description of the objects similarities. Usually the inverse of distance is considered as a relative definition for similarity measure, by taking large values for showing the low similarity and vice versa. In the time series analysis subject, close time series with the same shape and behavior are considered similar, regardless of having unequal samples and speed. Also, in this paper the distance function is defined as $D\left(X_1, X_2\right): n_1 \times n_2 \rightarrow \left[0, \infty\right)$ which can be included with following conditions [30]:

1) non-negativity $D\left(X_1, X_2\right) \geq 0$
2) identity of indiscernible $D\left(X_1, X_2\right) = 0 \Leftrightarrow X_1 = X_2$
3) symmetry $D\left(X_1, X_2\right) = D\left(X_2, X_1\right)$
4) triangle inequality $D\left(X_1, X_2\right) \leq D\left(X_1, X_2\right) + D\left(X_2, X_3\right)$

The first condition is inferred by the others. If (1) and (2) are satisfied, the distance function is positive-definite. Conditions (1), (2) and (3) together define symmetric function. If all of these conditions are satisfied, the function is considered to be a metric. The following properties should be existing for the desired distance function in the task of time series analysis:

- Measure the shape similarity of two time series
- Measure the physical closeness between two time series
- Compare time series with inconsistent temporal indexing

3. MULTIVARIATE TIME SERIES SIMILARITY MEASURES

In this section, common time series similarity measures developed in the literature are reviewed. The multivariate time series similarity measures compare overall shape of the time series by measuring closeness of time series. Similarity measures can be divided into four main categories as follows:

3. 1. Lock-Step Measures Methods in this category compare multivariate time series data samples one by one based on the temporal index. It means comparison of the i-th sample of one time series to the i-th sample of another. This kind of measures are limited to multivariate time series with equal length that is not applicable for most of the cases. ED [11, 31] and correlation are two famous lock-step similarity measures. Figure 1 shows the intuition behind Lock-step measures.

3. 2. Elastic Measures In the elastic measure category, the problem of aligning multivariate time series with different speed, different sampling rate, and inconsistent temporal scales is resolved by warping the temporal dimension. The basic idea of these methods is the Levenstein Distance (LD) [32], also known as edit distance, which is the smallest number of insertions, deletions, and substitutions needed to change one string to another. The elastic distance and warping path between two 2-d multivariate time series examples is given in Figure 2.

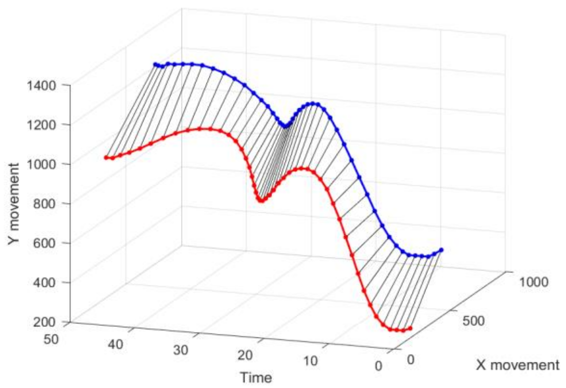

Figure 1. An illustration of a Lock-step measure (one-to-one mapping of MTS samples).

3. 2. 1. DTW The dynamic time warping, which shares many similarities with LD, was proposed in 1970 and 1971 to align multivariate time series with time shift tolerances [12, 13]. DTW distance applies local scaling of the temporal dimension. It guarantees to keep the order of multivariate time series samples and also it is sensitive to noise.

3. 2. 2. Constrained DTW Constrained DTW is one of the most useful variants of DTW to speed up and control deviation from the diagonal path (one-to-one matching). cDTW similarity measure constrains temporal scaling with Sakoe-Chiba Band [33], which consider a sliding window for temporal deviation. The size of sliding windows greatly affects quality of calculated similarity measure.

3. 2. 3. Weighted DTW Weighted DTW technique weighs each multivariate time series sample according to the temporal deviation [2]. Actually, it is a penalty weight that is proportionate to the warping difference and it is a soft version of cut-off cDTW.

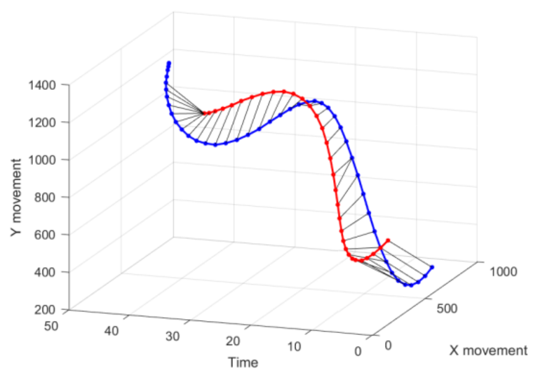

Figure 2. An example of an elastic measure (one-to-many mapping of MTS samples).

3. 2. 4. LCSS LCSS measures the longest common subsequence between multivariate time series based on the concept of edit distance [14]. Original LCSS measure is increased with the matching concept between two sequences. LCSS distance is robust against noise by using the threshold value on the distances between time series samples [25].

3. 2. 5. EDR EDR is another edit distance based similarity measure like LCSS that works by assigning a penalty to the gap between matched multivariate time series samples based on the length of the gap [15]. The method using a distance threshold to find a valid match between multivariate time series data samples.

3. 2. 6. ERP ERP is an edit-based measure that uses the merits of DTW and EDR, by considering a reference point for computing the distance where there is a gap in multivariate time series aligning [16]. The motivation for introducing the ERP is making EDR as a metric distance with a real penalty that defined by the distance to the reference point.

3. 2. 7. SWALE Morse and Patel proposed SWALE similarity measure based on edit distance that rewards matching and penalizes gaps [17]. In addition, the matching threshold is still used to find matching against noise.

3. 2. 8. TWED Marteau et al. [18] presented TWED metric distance that encompasses both LCSS and DTW characteristics. They redefine edit distance operations for measuring similarity. The originality of TWED lies in the way to control the stiffness, which is a multiplicative penalty that penalizes the deviation in the temporal dimension, unlike the constrained DTW that limits the deviation in the temporal dimension.

3. 2. 9. MSM Stefan et al. introduced the MSM metric that conceptually is an edit-based approach [19]. In this method, the similarity is estimated by presenting a set of new operations. Move, split and merge defined as three MSM operations with an associated cost. The Move operation is equivalent to substitute operation in the edit-based distance. Split and merge operations are different from insertion and deletion, however, it is achievable by combining MSM operations. The operation cost is not the same and they depend on the value of adjacent multivariate time series samples.

3. 3. Geometry-based Measures This kind of measures uses the shape as a geometric feature of the multivariate time series.

3. 3. 1. Hausdorff The Hausdorff distance shows the spatial similarity between two multivariate time series that measures how multivariate time series are far from each other [34]. If every sample of either multivariate time series is close to some other multivariate time series samples, then the Hausdorff distance will be low. Conventional Hausdorff distance considers not only the sampling point, but also every point of the sequences, thus it is a costly measure. The simple version of it, was proposed based on point to segment distance. Finally, the largest mismatched distance is selected that determines the value of unidirectional Hausdorff distance. Hausdorff can tolerate point disturbances, but it is sensitive to noise.

3. 3. 2. Frechet The Fréchet method considers the data samples with their orders along the continuous sequences [21]. Imagine a dog and dog's owner walking on two paths with keeping continuity from the start point to the end point. The shortest leash which is needed to connect the dog and its owner is the Fréchet distance between two traversed paths (time series). The free space is defined to simplify the computation of the Fréchet distance. That is a set of two time series points whose pair distance is less than a threshold. By finding the minimum value of the threshold, the Fréchet distance is achieved. The shortest distance that needs to connect is the Frechet distance between two multivariate time series. Eiter et al. [20] presented the discrete Fréchet (disFréchet) distance to approximate the exact Fréchet distance efficiently based on the recursive model. This method has reduced the complexity of discrete Fréchet distance.

3. 3. 3. SSPD The Symmetric Segment Path Distance (SSPD) is dependent on the point-to-segment distance like the Hausdorff [22]. The SSPD distance computes the minimum point-to-segment distance for every point of the first multivariate time series in all segments of the other one. Afterward, the average of the computed distance of every multivariate time series sample is reported as SSPD distance.

3. 4. Differentia-based MeasuresTW The first order difference of multivariate time series is the basis of similarity measures in this category.

3. 4. 1. DDTW Gorecki and Luczak introduced a derivative based distance that is a weighted combination of the DTW distance of raw multivariate time series with the first-order multivariate time series differences [23]. The weighting parameter could highly affect the method efficiency.

3. 4. 2. CIDTW Batista et al. [24] presented the CIDTW similarity measure as a weighting method to compensate the complexity difference of two comparing multivariate time series with the summation squares of the first-order difference. This weight is a multiplicative complexity-based value that weighs the DTW distance.

A summary of the mentioned similarity measures is shown in Table 1. The first column shows the category of similarity measures, the second column is the method name with references. Mathematic definitions are presented in the third column. In the fourth and fifth columns, distance type and time complexity are given, respectively.

TABLE 1. A summary of multivariate time series similarity measures

Method [Ref]	$D_{method}(X_1,X_2)$	Time Complexity
DTW [13]	$\begin{cases} 0 & \text{if } n_1=0 \text{ and } n_2=0 \\ \infty & \text{if } n_1=0 \text{ or } n_2=0 \\ d_{eucl}(x_{1,i},x_{2,j})+\min\begin{cases} D_{DTW}(rest(X_1),rest(X_2)) \\ D_{DTW}(rest(X_1),X_2) \\ D_{DTW}(X_1,rest(X_2)) \end{cases} & \text{otherwise} \end{cases}$	$O(n_1.n_2)$
WDTW [2]	$\begin{cases} 0 & \text{if } n_1=0 \text{ and } n_2=0 \\ \infty & \text{if } n_1=0 \text{ or } n_2=0 \\ w(x_{1,i},x_{2,j})+\min\begin{cases} D_{DTW}(rest(X_1),rest(X_2)) \\ D_{DTW}(rest(X_1),X_2) \\ D_{DTW}(X_1,rest(X_2)) \end{cases} & \text{otherwise} \end{cases}$, $w(x_{1,i},x_{2,j})=\frac{w_{max}}{1+e^{-g\times(\lvert i-j\rvert-m/2)}}\times d_{eucl}(x_{1,i},x_{2,j})$	$O(n_1.n_2)$
LCSS [14]	$1-\frac{LCSS(X_1,X_2)}{\min\{n_1,n_2\}}$, $LCSS(X_1,X_2)=\begin{cases} 0 & \text{if } n_1=0 \text{ or } n_2=0 \\ LCSS(rest(X_1),rest(X_2))+1 & \text{if } d_{eucl}(x_{1,i},x_{2,j})<\epsilon \\ \max\begin{cases} LCSS(rest(X_1),X_2) \\ LCSS(X_1,rest(X_2)) \end{cases} & \text{otherwise} \end{cases}$	$O(n_1.n_2)$
EDR [15]	$\begin{cases} n_2 & \text{if } n_1=0 \\ n_1 & \text{if } n_2=0 \\ \min\begin{cases} D_{EDR}(rest(X_1),rest(X_2))+subcost_{EDR}(x_{1,i},x_{2,j}) \\ D_{EDR}(rest(X_1),X_2)+1 \\ D_{EDR}(X_1,rest(X_2))+1 \end{cases} & \text{otherwise} \end{cases}$, $subcost_{EDR}(x_{1,i},x_{2,j})=\begin{cases} 0 & \text{if } d_{eucl}(x_{1,i},x_{2,j})<\epsilon \\ 1 & \text{otherwise} \end{cases}$	$O(n_1.n_2)$

ERP [16]	$\begin{cases} \sum_{j=1}^{n_2} d_{eucl}(gap, x_{2,j}) & \text{if } n_1 = 0 \\ \sum_{i=1}^{n_1} d_{eucl}(x_{1,i}, gap) & \text{if } n_2 = 0 \\ \min \begin{cases} D_{ERP}(rest(X_1), rest(X_2)) + d_{eucl}(x_{1,i}, x_{2,j}) \\ D_{ERP}(rest(X_1), X_2) + d_{eucl}(x_{1,i}, gap) \\ D_{ERP}(X_1, rest(X_2)) + d_{eucl}(gap, x_{2,j}) \end{cases} & \text{otherwise} \end{cases}$	$O(n_1.n_2)$
SWALE [17]	$\begin{cases} n_2 \times g_c & \text{if } n_1 = 0 \\ n_1 \times g_c & \text{if } n_2 = 0 \\ D_{SWALE}(rest(X_1), rest(X_2)) + r_m & \text{if } d_{eucl}(x_{1,i}, x_{2,j}) < \varepsilon \\ \max \begin{cases} D_{SWALE}(rest(X_1), X_2) + g_c \\ D_{SWALE}(X_1, rest(X_2)) + g_c \end{cases} & \text{otherwise} \end{cases}$	$O(n_1.n_2)$
TWED [18]	$\begin{cases} 0 & \text{if } n_1 = 0 \text{ and } n_2 = 0 \\ \infty & \text{if } n_1 = 0 \\ \infty & n_2 = 0 \\ \min \begin{cases} D_{TWED}(rest(X_1), rest(X_2)) + cost_{match}(x_{1,i}, x_{2,j}) \\ D_{TWED}(rest(X_1), X_2) + cost_{delete}(x_{1,i}, x_{1,i-1}) \\ D_{TWED}(X_1, rest(X_2)) + cost_{delete}(x_{2,j}, x_{2,j-1}) \end{cases} & \text{otherwise} \end{cases}$, $cost_{match}(x_{1,i}, x_{2,j}) = 2 * \nu * \lvert i - j \rvert + d_{eucl}(x_{1,i}, x_{2,j}) + d_{eucl}(x_{1,i-1}, x_{2,j-1})$ $cost_{delete}(x_{k,i}, x_{k,i-1}) = \nu + \lambda + d_{eucl}(x_{k,i}, x_{k,j})$	$O(n_1.n_2)$
MSM [19]	$\begin{cases} d_{eucl}(x_{1,1}, x_{2,1}) & \text{if } n_1 = 1 \text{ and } n_2 = 1 \\ D_{MSM}(rest(X_1), X_2) + cost_{MSM}(x_{1,i}, x_{1,i-1}, x_{2,1}) & \text{if } n_1 \neq 1 \text{ and } n_2 = 1 \\ D_{MSM}(X_1, rest(X_2)) + cost_{MSM}(x_{2,j}, x_{1,1}, x_{2,j-1}) & \text{if } n_1 = 1 \text{ and } n_2 \neq 1 \\ \min \begin{cases} D_{MSM}(rest(X_1), rest(X_2)) + d_{eucl}(x_{1,i}, x_{2,j}) \\ D_{MSM}(rest(X_1), X_2) + cost_{MSM}(x_{1,i}, x_{1,i-1}, x_{2,j}) \\ D_{MSM}(X_1, rest(X_2)) + cost_{MSM}(x_{2,j}, x_{1,i}, x_{2,j-1}) \end{cases} & \text{otherwise} \end{cases}$, $cost_{MSM}(x_1, x_2, x_3, c) = \begin{cases} c & \text{if } x_2 \le x_1 \le x_3 \text{ or } x_2 \ge x_1 \ge x_3 \\ c + \min \begin{cases} d_{eucl}(x_1, x_2) \\ d_{eucl}(x_1, x_3) \end{cases} & \text{otherwise} \end{cases}$	$O(n_1.n_2)$
Hausdorff [34]	$\max \begin{cases} \max_{x_{1,i} \in X_1} \left\{ \min_{xs_{2,j} \in Xs_2} \left\{ d_{p2s}(x_{1,i}, xs_{2,j}) \right\} \right\} \\ \max_{x_{2,i} \in X_2} \left\{ \min_{xs_{1,j} \in Xs_1} \left\{ d_{p2s}(x_{2,i}, xs_{2,j}) \right\} \right\} \end{cases}$	$O(n_1.n_2)$
disFréchet [20]	$\begin{cases} d_{eucl}(x_{1,1}, x_{2,1}) & \text{if } n_1 = 1 \text{ and } n_2 = 1 \\ \max \begin{cases} D_{disFréchet}(rest(X_1), X_2) \\ d_{eucl}(x_{1,i}, x_{2,1}) \end{cases} & \text{if } n_1 > 1 \text{ and } n_2 = 1 \\ \max \begin{cases} D_{disFréchet}(X_1, rest(X_2)) \\ d_{eucl}(x_{1,1}, x_{2,j}) \end{cases} & \text{if } n_1 = 1 \text{ and } n_2 > 1 \\ \max \begin{cases} \min \begin{cases} D_{disFréchet}(rest(X_1), rest(X_2)) \\ D_{disFréchet}(rest(X_1), X_2) \\ D_{disFréchet}(X_1, rest(X_2)) \end{cases} \\ d_{eucl}(x_{1,i}, x_{2,j}) \end{cases} & \text{otherwise} \end{cases}$	$O(n_1.n_2)$
SSPD [22]	$\frac{1}{2n_1} \sum_{i=1}^{n_1} \min_{j \in [1,\ldots,n_2-1]} d_{p2s}(x_{1,i}, xs_{2,j}) + \frac{1}{2n_2} \sum_{i=1}^{n_2} \min_{j \in [1,\ldots,n_1-1]} d_{p2s}(x_{2,i}, xs_{1,j})$	$O(n_1.n_2)$
DDTW [23]	$\cos\alpha \times D_{DTW}(X_1, X_2) + \sin\alpha \times D_{DTW}(diff(X_1), diff(X_2))$	$O(n_1.n_2)$
CIDTW [24]	$D_{DTW}(X_1, X_2) \times \frac{\max \begin{cases} complexity(X_1) \\ complexity(X_2) \end{cases}}{\min \begin{cases} complexity(X_1) \\ complexity(X_2) \end{cases}}$, $complexity(X) = \sqrt{\Sigma (diff(X))^2}$	$O(n_1.n_2)$

4. EVALUATION FRAMEWORK

4. 1. Computation Time The time required to compute the distance matrix for all multivariate time series datasets is calculated as a criterion to compare the computation cost between different similarity measures.

4. 2. Clustering Scheme The selected clustering techniques and the clustering evaluation are examined in this section. We will study different clustering methods obtained with the same algorithm but with mentioned measures to evaluate considered similarity measures. The choice of clustering technique is limited by the characteristics of multivariate time series data. K-means algorithm and spectral clustering cannot be used for multivariate time series data [22]. Dbscan and k-medoid clustering can be used, but they are not efficient. As a matter of fact, these algorithms are based on the nearest neighbor and required to be metric [22]. Most of the studied similarity measures are not metrics.

To cluster the multivariate time series datasets, we will focus on hierarchical cluster analysis (HCA). Indeed, the HCA does not need the metric similarity measure. Also, the HCA does not need any extra parameters and only need the similarity matrix, thus it can cluster multivariate time series with different length. The agglomerative HCA with seven different linkage options was used in our experiments [35, 36].

A clustering algorithm aims to group a set of objects in such a way that objects in the same group are more similar to each other than to those in other clusters. In order to evaluate the quality of clustering F -measure criterion was used. Since the labels returned by a clustering run are arbitrary and the ground-truth label for datasets is available, the F -measure criterion was used as follows [37]:

$$\mathrm{F}\left(C,C^{*}\right)=\frac{1}{N}\sum_{i=1}^{k_C}N_{C_i}\max_{j=1,\ldots,k_{C^*}}\left\{\frac{2\left|C_i\cap C_j^{*}\right|}{\left|C_i\right|+\left|C_j^{*}\right|}\right\} \quad (6)$$

where C_i is the i-th ground truth class, C_j^* is the j-th cluster, N is the number of multivariate time series in dataset, N_{C_i} is the number of multivariate time series in C_i, k_C is the number of ground truth class in C and k_{C^*} is the number of clusters in C^*.

In this study, k_C and k_{C^*} are assumed to be always equal. As the characteristics of used multivariate time series datasets, that will be discussed in section 4.4, the number of clusters is known and will be used directly.

4. 3. Parameter Tuning Several investigated similarity measures have one or more controlling parameters that choosing these parameters directly affects the measures productivity.

TABLE 2. Parameter grid for the considered similarity measures (recall that n_1 and n_2 corresponds to the length of the input multivariate time series)

Measure	Parameter	Min	Max	Steps
cDTW	Windows - δ	$\lvert n_1-n_2\rvert$	$25\%\max(n_1,n_2)$	5
WDTW	Curvature - g	0	1	5
LCSS	Threshold - ε	$2\% std(x_{X_k})$	$std(x_{X_k})$	5
EDR	Threshold - ε	$2\% std(x_{X_k})$	$std(x_{X_k})$	5
ERP	Penalty - *gap*	0	$\mp 3.std(x_{X_k})$	3
SWALE	Reward - r_m	$50\ std(x_{X_k})$		-
SWALE	Penalty - g_c	0	r_m	5
SWALE	Threshold – ε	$2\% std(x_{X_k})$	$std(x_{X_k})$	5
TWED	Stiffness - ν	10^{-5}	10^{0}	5
TWED	Penalty - λ	0	$std(x_{X_k})$	5
MSM	Cost - c .	$\{10^{-2},10^{-1},10^{0},10^{1},10^{2},10^{3},10^{4},10^{5}\}$		
DDTW	Ratio - α	1	$\pi/2$	5

In this experiment, the grid search within a suitable range of parameters was used that can be chosen according to the given specification in the introducing papers of each measure, as described in Table 2.

For each similarity measure, we analyze the F -measure for 14 different clustering variations. The parameter with the best F -measure for the highest number of clustering variations was s selected as the similarity measure parameter and the other similarity measure parameters were discarded. Finally, the selected parameter was used to evaluate final F -measure.

4. 4. Datasets Experiments were performed using 23 publicly-available labeled multivariate time series datasets with varying properties that are presented briefly in Table 3. They include synthetic, as well as real-world datasets. There are some criteria to characterize the datasets such as the number of time series, average length of time series and average shape complexity [38] as mentioned in Table 3.

The shape complexity can be calculated as follows:

$$\xi_{X_k}=\frac{d_{eucl}\left(x_{k,1},x_{k,n_k}\right)}{\sum_i l_{xs_{k,i}}} \quad (7)$$

where, ξ_{X_k} is the shape complexity of time series X_k and $l_{xs_{k,i}}$ is the length of multivariate time series segment $xs_{k,i}$.

4. 4. Statistical Significance The non-parametric rank-based test is an accepted statistic for comparing the performance of n_{cl} clustering with different similarity measures over n_X datasets [39, 40]. A null hypothesis assumes that the average performance rank of n_{cl} similarity measures on n_X multivariate time series datasets are the same (not significantly different). There is an alternative hypothesis against the null hypothesis which assumes at least one measure's mean rank is different. The M is a n_X by n_{cl} matrix that includes the F -measure value of clustering results. At the first stage, the performance rank of each similarity measure was evaluated for each dataset separately and make the matrix R, where the r_{ij} element shows the rank of the j^{th} similarity measure on the i^{th} dataset. The rank of measures with the equal F -measure were averaged. The average rank of each measure was denoted as $R_j = \frac{1}{n_X}\sum_i r_{ij}$. Under the null hypothesis, the ranks of all similarity measures were equal, the Friedman statistics [48] F_F can be approximated by F-distribution with $(n_{cl}-1)$ and $(n_{cl}-1)(n_X-1)$ degree of freedom as follow:

$$F_F = \frac{(n_X-1)\chi_F^2}{n_X(n_{cl}-1)-\chi_F^2}$$
$$\chi_F^2 = \frac{12.n_X}{n_{cl}(n_{cl}+1)}\left[\sum_j R_j^2 - \frac{n_{cl}(n_{cl}+1)^2}{4}\right] \quad (8)$$

where F_F is the Friedman statistics value.

If the null hypothesis is rejected based on the test results, the further family-wise comparisons will be needed. The Holland [49] post-hoc method was used to compensate multiple family-wise comparisons.

TABLE 3. Datasets characterization

Dataset [ref]	Size	#Class	ξ_{X_k}	Source	Type
ASL-10 [14]	699	10	0.03	Australian Sign Language	Real
ASL-35[41]	700	35	0.02		Real
VMT [42]	1500	15	0.87	Vehicle	Real
SM [42]	2500	50	0.25	Vehicle	Synthetic
CROSS [43]	1900	19	0.81	Vehicle	Real
I5 [43]	806	8	1.00	Vehicle	Real
I5SIM1 [43]	800	8	0.81	Vehicle	Synthetic
I5SIM2 [43]	1600	8	0.60	Vehicle	Synthetic
I5SIM3 [43]	1600	16	0.60	Vehicle	Synthetic
LABOMNI [43]	209	15	0.49	Human	Real
FT [44]	3102	2	0.61	Fish	Real
HC-digit [45]	198	9	0.47	Hand written	Real
HC [45]	1363	35	0.47		Real
CAL1 [46]	400	2	0.88	Human	Synthetic
CAL2 [46]	670	3	0.88	Human	Synthetic
CAL3 [46]	900	4	0.88	Human	Synthetic
CAL4 [46]	1210	5	0.79	Human	Synthetic
CAL5 [46]	1130	8	0.75	Human	Synthetic
CAL6 [46]	380	3	0.76	Human	Synthetic
CAL7 [46]	220	3	0.95	Human	Synthetic
CAL8 [46]	120	18	0.85	Human	Synthetic
CAL9 [46]	300	4	0.66	Human	Synthetic
SIGNATURE [47]	1600	40	0.20	Human	Real

5. RESULTS

In this section, the effectiveness of 14 similarity measures include: DTW, cDTW, WDTW, LCSS, EDR, ERP, SWALE, TWED, MSM, Hausdorff, disFréchet, SSPD, DDTW, and the CIDTW are evaluated over 23 publicly-available datasets. The computational time and the results of the clustering technique using each of mentioned similarity measures are compared.

All distances have been implemented in MATLAB and Mex and are available in the time series analysis package available on github[2]. The entire simulation was conducted on a CORE-I7 computer with 16GB of RAM running for over a month.

5. 1. Time Complexity The time required to compute the distance matrix for all multivariate time series datasets was calculated as a criterion to compare the computation cost between different similarity measures. In Figure 3, the total time needed to compute the considered similarity measures for all multivariate time series datasets is shown.

It should be mentioned that all considered similarity measures, except cDTW technique, run in $O(n_1.n_2)$ but this is slightly confounded when considering parameter options. WDTW is the similarity measure that requires the most computation time. The cDTW measure shows the lowest total time, it is predictable because it has a lower complexity than other competitors.

[2] https://github.com/amirsalarpour/Time-Series-Similarity

TABLE 4. The F -measure for all considered similarity measures and multivariate time series datasets. The last row is the average rank of each measure across all datasets. The best performances are bolded

Dataset	Similarity measures													
	DTW	**cDTW**	**WDTW**	**LCSS**	**EDR**	**ERP**	**SWALE**	**TWED**	**MSM**	**Hausdorff**	**disFréchet**	**SSPD**	**DDTW**	**CIDTW**
ASL-10	0.946	0.946	0.947	0.944	0.947	0.946	**0.947**	0.947	0.919	0.900	0.830	0.911	0.889	0.947
ASL-35	0.938	**0.962**	0.953	0.914	0.943	0.930	**0.962**	0.943	0.927	0.812	0.863	0.927	0.925	0.937
VMT	0.950	0.950	**0.963**	0.942	0.910	0.882	0.956	0.940	0.930	0.950	0.950	0.945	0.961	0.951
SM	0.979	0.978	0.979	**0.979**	0.972	0.962	0.962	0.977	0.976	0.969	0.955	0.977	0.978	0.974
CROSS	0.866	0.866	0.869	0.973	0.832	0.816	0.970	**0.995**	0.971	0.994	0.994	0.991	0.867	0.869
I5	0.807	0.863	0.807	0.953	0.936	0.834	0.943	0.906	0.900	0.717	0.730	**0.982**	0.863	0.902
I5SIM1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.995	**1.000**
I5SIM2	0.895	0.895	0.895	**1.000**	0.940	0.711	**1.000**	0.895	0.895	**1.000**	**1.000**	**1.000**	0.895	0.895
I5SIM3	0.906	0.906	0.906	0.999	0.924	0.916	**1.000**	0.899	0.954	0.856	0.835	0.855	0.936	0.906
LABOMNI	0.880	0.880	0.880	0.879	0.820	0.824	0.921	0.846	0.868	0.860	**0.954**	0.863	0.886	0.880
FT	**0.995**	**0.995**	**0.995**	**0.995**	0.938	**0.995**	**0.995**	0.924	0.908	0.991	0.991	**0.995**	0.991	0.991
HC-digit	0.842	0.876	0.893	0.927	0.884	0.905	0.927	**0.930**	0.925	0.898	0.916	0.916	0.894	0.925
HC	0.920	0.931	0.920	0.974	**0.974**	0.922	0.974	0.936	0.921	0.949	0.958	0.878	0.941	0.965
CAL1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
CAL2	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
CAL3	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
CAL4	0.924	0.927	0.919	0.972	0.963	0.924	0.934	0.943	0.938	0.895	0.957	0.909	0.927	**1.000**
CAL5	0.944	0.944	0.944	0.930	0.941	0.944	0.964	**0.965**	0.962	0.940	0.950	0.950	0.950	0.946
CAL6	**0.996**	**0.996**	**0.996**	0.984	0.984	**0.996**	0.981	0.984	0.984	0.974	0.974	0.982	0.982	**0.996**
CAL7	0.920	0.920	0.920	0.931	0.931	0.920	0.931	0.920	0.923	0.901	0.926	0.920	0.920	**0.984**
CAL8	0.771	0.771	0.771	0.900	0.906	0.771	**0.924**	0.873	0.859	0.800	0.802	0.748	0.771	0.854
CAL9	0.955	0.955	0.955	0.943	0.955	0.955	0.955	0.955	0.955	0.955	0.955	0.955	0.955	**0.996**
SIGNATURE	**0.976**	0.957	0.970	0.954	0.968	0.908	0.974	0.968	0.953	0.941	0.954	0.956	0.958	0.954
Average Rank	8.09	7.48	7.17	6.26	7.24	9.52	**4.54**	6.65	8.15	9.67	7.93	8.13	8.11	6.04

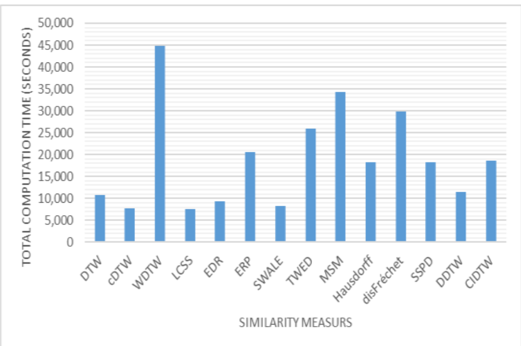

Figure 3. Total computation time in seconds for all considered similarity measures

DTW, LCSS, EDR, and SWALE measures all having the same order of time complexity, are the fastest computed methods after cDTW. All differential based and elastic measures need to compute the ED between a pair of multivariate time series samples and the only difference is their cost function. Hausdorff and SSPD measures were also computed in the same way, hence it explains why they have almost the same computation time.

5. 2. Analysis of Clustering

The performances of considered similarity measures on each multivariate time series dataset are presented in Table 4. Every column includes the F -measure value of the best clustering variant among different similarity measure parameters and different HCA options. The best performance over each multivariate time series dataset was bolded. The last row contains the average rank of each measure across all datasets that is the average position after sorting the F -measure for a given dataset in descending order.

It is observable that the SWALE perform as the best similarity measure based on average rank. It should be noted, the SWALE technique uses three different contrlling techniques in a same measure to produce the similairty. It gains from collaborating the distance threshold, penalty and reward options in calculating the similarity and need three parameters to set. Although, the SWALE technique has more parameters, that need to be optimised, but if the proper values are chosen, it colud be well similarity measure between time series.

Figure 4 presents the behavior of SWALE, as the best measure based on average rank, versus each competitor based on the number of datasets, where SWALE produces respectively better performance, equal performance, and worse performance compared to each of them. The goal of this experiment was to compare the competitive performance of SWALE based on clustering compared to other similarity measures. As can be seen from Figure 4, the LCSS measure has the lowest number of datasets that works worse than SWALE. On the opposite side, the TWED technique has the highest number of datasets where its performance is better than SWALE.

To provide a more intuitive illustration of the performance of the similarity measures compared to the SWALE as the best performer on our datasets based on the average rank, the pairwise comparison conducted through the scatter plot was used. In a scatter plot, the F -measure of the SWALE was used as the y coordinate of a dot and the F -measure of the similarity measure under comparison was used as x coordinates of a dot, where each dot represents a particular dataset.

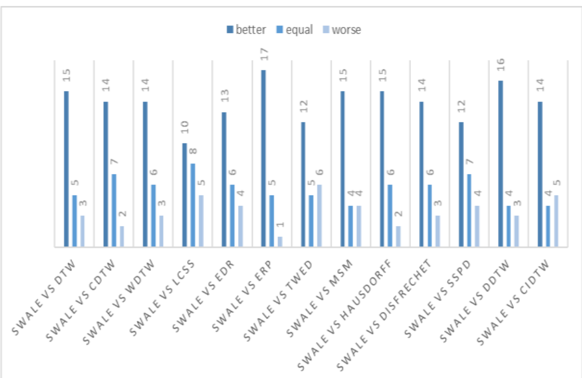

Figure 4. The number of datasets which SWALE produced better, equal, or worse performance compared to other similarity measures

Each scatter plot has the label "SWALE versus A", a plot above the line indicates that SWALE is a better performer than A. The further a dot is from the line, the greater the margin of F -measure improvement.

The more dots on one side of the line indicates that the better one similarity measure is compared to the other. The performance of SWALE against its elastic competitors is shown in Figure 5. It can be observed in Figure 5 (c) that the effectiveness of SWALE is slightly better than that of LCSS.

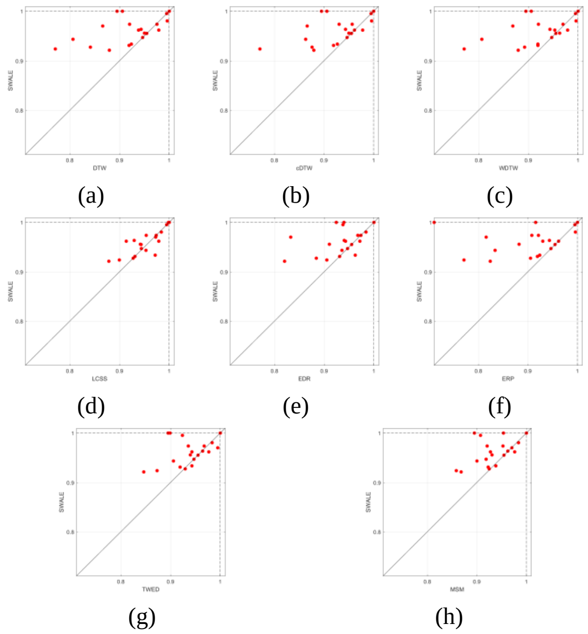

Figure 5. Pairwise comparison of SWALE against elastic similarity measures. (a) SWALE vs DTW, (b) SWALE vs cDTW, (c) SWALE vs WDTW, (d) SWALE vs LCSS, (e) SWALE vs EDR, (f) SWALE vs ERP, (g) SWALE vs TWED, and (h) SWALE vs MSM

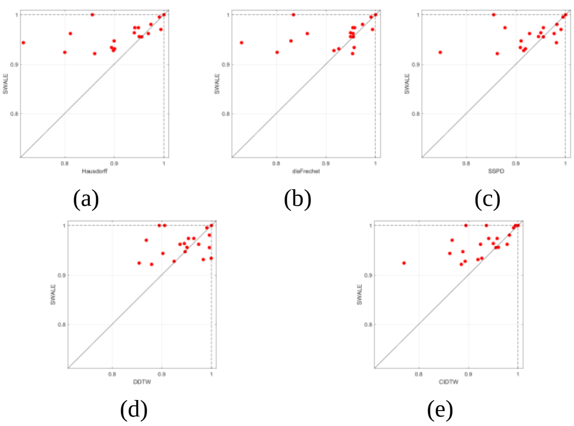

Figure 6. Pairwise comparison of SWALE against geometry-based and differential-based similarity measures. (a) SWALE vs Hausdorff, (b) SWALE vs disFréchet, (c) SWALE vs SSPD, (d) SWALE vs DDTW, and (e) SWALE vs CIDTW

As can be seen in Figure 5, except part (c), the SWALE measure is clearly superior over DTW, cDTW, EDR, ERP, TWED, and MSM measures.

Figure 6 depicts the performance of SWALE against its geometry-based and differential-based competitors. As shown in Figure 6 (a)-(c) SWALE measure is clearly superior to Hausdorff, disFréchet, and SSPD similarity measures on tested datasets.

It can be observed in Figure 6 (c) that the effectiveness of SWALE is slightly better than that of LCSS. As can be seen from Figure 6, except the part (c), the SWALE measure is clearly superior to DTW, cDTW, EDR, ERP, TWED, and MSM measures. From Figure 6 (d), it can be seen that SWALE is a better performer than DDTW measure. Figure 6 (e) shows SWALE measure largely outperforms the CIDTW meaure.

5. 3. Analysis of Statistical Significance The Friedman (Iman-Davenport) test results, with having 14 similarity measures and 23 datasets, is equal to 2.5743 according to F-distribution. Hence, the null hypothesis is rejected based on the F-distribution with 13 and 13*22 degree of freedom and with p-value 0.0022 at a high confidence level.

Demsar [39] recommends grouping classifiers into cliques, within which there is no significant difference in rank. This allows the average ranks and groups of not significantly different classifiers to be plotted on an order line in a graph referred to as a critical difference diagram. In this way, Figure 7 shows the critical statistical difference diagram for 14 similarity measures over the 23 datasets.

As shown in Figure 7 there is no similarity measure that significantly outperforms the others. There are four cliques within which no significant difference is observed. The top clique contains all but ERP and Hausdorff. It means there is no significantly difference between similarity measures (other than ERP and Hausdorff) for the clustering task based on our multivariate time series datasets. These results do not lend any support to each of similarity measure over other multivariate time series similarity measures.

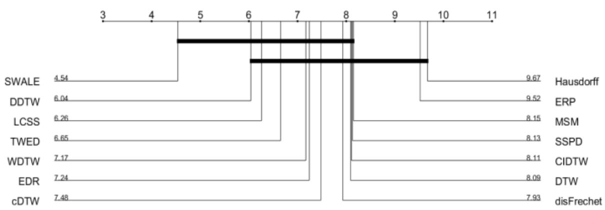

Figure 7. The average ranks for 14 similarity measures over 23 datasets

5. CONCLUSIONS

In this paper, 14 similarity measures on 23 publicly available datasets with different characteristics for multivariate time series data were extensively evaluated. The computation time and clustering performance were evaluated and discussed in detail. Clustering performance was assessed in terms of F -measure and statistical significance. Our main findings are as follows:

- The WDTW and cDTW methods spend the most and the least computation time, respectively.
- Between geometry-based similarity measures, the Hausdorff distance shows the lowest computation time and disFréchet illustrates the highest time complexity.
- The SWALE measure, was originally proposed by Morse and Patel, consistently performs better than all considered measures.
- The SWALE technique obtained the best average rank among all similarity methods (4.54), followed by TWED with an average rank of 6.04. Also, Hausdorff distance showed the worst F -measure in nearly all datasets and showed the weakest average rank (9.67).
- Finally, we conclude that there was no specific similarity measure that statistically significantly outperformed the other techniques based on our datasets.

The large-scale-based experimental evaluation of multiple approaches is necessary for any mature research field, because it opens up your view to select the most appropriate one. Besides getting an idea to use relevant similarity measure, it provides the unified framework to compare and analyze multivariate time series data. Adding alternative measures and using more datasets may lead to more comprehensive results. Also, the application of similarity measures for a specific goal, such as pattern extraction, prediction, vehicle analysis could be investigated.

6. REFERENCES

1. Keogh, E., Xi, X., Wei, L. and Ratanamahatana, C. A., "The UCR time series classification/clustering homepage", ***URL= http://www. cs. ucr. edu/eamonn/time series data*** Vol., (2006)
2. Jeong, Y.-S., Jeong, M. K. and Omitaomu, O. A., "Weighted dynamic time warping for time series classification", ***Pattern Recognition*** Vol. 9, (2011), 2231-2240.
3. Kantz, H. and Schreiber, T., "Nonlinear time series analysis", Cambridge university press, (2004).
4. Fu, Z., Hu, W. and Tan, T., "Similarity based vehicle trajectory clustering and anomaly detection", Image Processing, 2005. ICIP 2005. IEEE International Conference on, IEEE, (2005); Liao, T. W., "Clustering of time series data—a survey", ***Pattern Recognition*** Vol. 11, (2005), 1857-1874.

5. Doroudyan, M., Owlia, M., Sadeghi, H. and Amiri, A., "Monitoring Financial Processes with ARMA-GARCH Model Based on Shewhart Control Chart (Case Study: Tehran Stock Exchange)", ***International Journal of Engineering-Transactions B: Applications*** Vol. 2, (2017), 270-278.

6. Shafiee-Chafi, M. and Gholizade-Narm, H., "A novel fuzzy based method for heart rate variability prediction", ***International Journal of Engineering-Transactions A: Basics*** Vol. 7, (2014), 1041-1049.

7. Aghabozorgi, S., Shirkhorshidi, A. S. and Wah, T. Y., "Time-series clustering–A decade review", ***Information Systems*** Vol., (2015), 16-38.

8. Kleist, C., "Time Series Data Mining Methods: A Review", Humboldt-Universität zu Berlin, (2015).

9. Darvishi, A. and Hassanpour, H., "A geometric view of similarity measures in data mining", ***International Journal of Engineering-Transactions C: Aspects*** Vol. 12, (2015), 1728-1734.

10. Faloutsos, C., Ranganathan, M. and Manolopoulos, Y., "Fast subsequence matching in time-series databases", ACM, (1994).

11. Rabiner, L. R. and Juang, B.-H., "Fundamentals of speech recognition", Vol., (1993)

12. Berndt, D. J. and Clifford, J., "Using dynamic time warping to find patterns in time series", KDD workshop, Seattle, WA, (1994).

13. Vlachos, M., Kollios, G. and Gunopulos, D., "Discovering similar multidimensional trajectories", Data Engineering, 2002. Proceedings. 18th International Conference on, IEEE, (2002).

14. Chen, L. and Ng, R., "On the marriage of lp-norms and edit distance", Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, VLDB Endowment, (2004).

15. Chen, L., Özsu, M. T. and Oria, V., "Robust and fast similarity search for moving object trajectories", Proceedings of the 2005 ACM SIGMOD international conference on Management of data, ACM, (2005).

16. Morse, M. D. and Patel, J. M., "An efficient and accurate method for evaluating time series similarity", Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM, (2007).

17. Marteau, P. F., "Time warp edit distance with stiffness adjustment for time series matching", ***IEEE Trans Pattern Anal Mach Intell*** Vol. 2, (2009), 306-318.

18. Stefan, A., Athitsos, V. and Das, G., "The move-split-merge metric for time series", ***IEEE transactions on Knowledge and Data Engineering*** Vol. 6, (2013), 1425-1438.

19. Eiter, T. and Mannila, H., "Computing discrete Fréchet distance" (1994).

20. Alt, H. and Godau, M., "Computing the Fréchet distance between two polygonal curves", ***International Journal of Computational Geometry & Applications*** Vol. 01n02, (1995), 75-91.

21. Besse, P. C., Guillouet, B., Loubes, J.-M. and Royer, F., "Review and Perspective for Distance-Based Clustering of Vehicle Trajectories", ***IEEE Transactions on Intelligent Transportation Systems*** Vol. 11, (2016), 3306-3317.

22. Górecki, T. and Łuczak, M., "Using derivatives in time series classification", ***Data Mining and Knowledge Discovery*** Vol., (2013), 1-22.

23. Batista, G. E., Keogh, E. J., Tataw, O. M. and De Souza, V. M., "CID: an efficient complexity-invariant distance for time series", ***Data Mining and Knowledge Discovery*** Vol. 3, (2014), 634-669.

24. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X. and Keogh, E., "Querying and mining of time series data: experimental comparison of representations and distance measures", ***Proceedings of the VLDB Endowment*** Vol. 2, (2008), 1542-1552.

25. Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P. and Keogh, E., "Experimental comparison of representation methods and distance measures for time series data", ***Data Mining and Knowledge Discovery*** Vol., (2013), 1-35; Serra, J. and Arcos, J. L., "An empirical evaluation of similarity measures for time series classification", ***Knowledge-Based Systems*** Vol., (2014), 305-314; Lines, J. and Bagnall, A., "Time series classification with ensembles of elastic distance measures", ***Data Mining and Knowledge Discovery*** Vol. 3, (2015), 565-592; Bagnall, A., Bostrom, A., Large, J. and Lines, J., "The great time series classification bake off: an experimental evaluation of recently proposed algorithms", ***Extended Version. CoRR, abs/1602.01711*** Vol., (2016)

26. Deza, M. M. and Deza, E., "Encyclopedia of distances", Springer, (2009).

27. Wei, W. W.-S., "Time series analysis", Addison-Wesley publ Reading, (1994).

28. Levenshtein, V. I., "Binary codes capable of correcting deletions, insertions, and reversals", Soviet physics doklady, (1966).

29. Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition", ***IEEE transactions on acoustics, speech, and signal processing*** Vol. 1, (1978), 43-49.

30. Hausdorff, F., "Mengenlehre", Walter de Gruyter Berlin, (1927).

31. Mori, U., Mendiburu, A. and Lozano, J. A., "Similarity measure selection for clustering time series databases", ***IEEE Transactions on Knowledge and Data Engineering*** Vol. 1, (2016), 181-195; Yuan, G., Sun, P., Zhao, J., Li, D. and Wang, C., "A review of moving object trajectory clustering algorithms", ***Artificial Intelligence Review*** Vol. 1, (2017), 123-144.

32. Wagner, S. and Wagner, D., "Comparing clusterings: an overview", Universität Karlsruhe, Fakultät für Informatik Karlsruhe, (2007).

33. Zhang, Z., Huang, K. and Tan, T., "Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes", Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, IEEE, (2006).

34. Kadous, W., "GRASP: Recognition of Australian sign language using Instrumented gloves", Vol., (1995)

35. Hu, W., Li, X., Tian, G., Maybank, S. and Zhang, Z., "An incremental DPMM-based method for trajectory clustering, modeling, and retrieval", ***IEEE Trans Pattern Anal Mach Intell*** Vol. 5, (2013), 1051-1065.

36. Morris, B. and Trivedi, M., "Learning trajectory patterns by clustering: Experimental studies and comparative evaluation", Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, (2009).

37. Beyan, C. and Fisher, R. B., "Detection of Abnormal Fish Trajectories Using a Clustering Based Hierarchical Classifier", BMVC, (2013).

38. Ramos-Garijo, R., Martín, S., Marzal, A., Prat, F., Vilar, J. M. and Llorens, D., "An input panel and recognition engine for on-line handwritten text recognition", ***Frontiers in Artificial Intelligence and Applications*** Vol., (2007), 223-230.

39. Calderara, S., Prati, A. and Cucchiara, R., "Mixtures of von mises distributions for people trajectory shape analysis", ***IEEE***

Transactions on Circuits and Systems for Video Technology Vol. 4, (2011), 457-471.

40. Yeung, D.-Y., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T. and Rigoll, G., "SVC2004: First international signature verification competition", Springer, (2004).
41. Demšar, J., "Statistical comparisons of classifiers over multiple data sets", ***Journal of Machine Learning Research*** Vol. Jan, (2006), 1-30.
42. García, S., Fernández, A., Luengo, J. and Herrera, F., "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power", ***Information Sciences*** Vol. 10, (2010), 2044-2064.
43. Friedman, M., "A comparison of alternative tests of significance for the problem of m rankings", ***The Annals of Mathematical Statistics*** Vol. 1, (1940), 86-92.
44. Holland, B. S. and Copenhaver, M. D., "An improved sequentially rejective Bonferroni test procedure", ***Biometrics*** Vol., (1987), 417-423.

An Empirical Comparison of Distance Measures for Multivariate Time Series Clustering

A. Salarpour, H. Khotanlou

RIV Lab, Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran

PAPER INFO

Paper history:
Received 15 October 2017
Received in revised form 06 December 2017
Accepted 21 December 2017

Keywords:
Multivariate Time Series
Similarity Measures
Clustering
Evaluation

چکیده

داده‌های سری‌های زمانی چند متغیره در زندگی روزمره و زمینه‌های متفاوتی از علوم به وفور یافت می‌شود و چگونگی اندازه‌گیری شباهت بین این نوع سری‌های زمانی یکی از بخش‌های اصلی روند آنالیز سری‌های زمانی است. حجم زیادی از تحقیقات انجام شده در این زمینه بر روی ارایه معیار های شباهت جدید متمرکز شده است. با وجود ارائه روش‌های بسیار برای محاسبه شباهت بین سری‌های زمانی، این زمینه همچنان از فقدان یک مطالعه مقایسه‌ای با استفاده از ارزیابی کمی در مقیاس مناسب رنج می‌برد. بدین منظور و برای فراهم کردن یک سنجش مقایسه ای، ارزیابی گسترده ای بر روی معیار های شباهت برای خوشه بندی سری‌های زمانی انجام داده‌ایم. بدین منظور، ۱۴ معیار شباهت معتبر (به همراه تنوعات آنها) و بررسی اثرگذاری آنها بر روی ۲۳ دیتاست سری زمانی چند متغیره (مربوط به کاربردهای مختلف) انجام شده است. در این مقاله، مقایسه تجربی در خصوص اثر بخشی معیارهای شباهت مبتنی بر خوشه‌بندی سلسه مراتبی ارائه شده است. علاوه بر این، آزمون آماری معناداری نیز برای سنجش معناداری بین بازدهی معیارهای شباهت، بکار گرفته شد. نتایج حاصل نشان داد که معیارهای شباهت بررسی شده بر اساس نتایج خوشه بندی عملکردی یکسان داشته و از لحاظ آماری و مبتنی بر آزمون غیرپارامتری اختلاف معناداری ندارند. نتایج بدست آمده یک دید مقایسه ای بین معیارهای شباهت و برای یافتن روش مناسب بر اساس بازدهی و پیچیدگی زمانی فراهم می کند.

doi*: 10.5829/ije.2018.31.02b.08*