Frobenius correlation based u-shapelets discovery for time series clustering

Vanel Steve Siyou Fotso, Engelbert Mephu Nguifo, Philippe Vaslin

Frobenius correlation based u-shapelets discovery for time series clustering

Vanel Steve SIYOU FOTSO[1], **Engelbert MEPHU NGUIFO**[1], **Philippe VASLIN**[1],
[1] University Clermont Auvergne, CNRS, LIMOS, F-63000 Clermont-Ferrand, France
siyou@isima.fr, mephu@isima.fr, vaslin@isima.fr

Abstract

An u-shapelet is a sub-sequence of a time series used for clustering a time series dataset. The purpose of this paper is to discover u-shapelets on uncertain time series. To achieve this goal, we propose a dissimilarity score called FOTS whose computation is based on the eigenvector decomposition and the comparison of the autocorrelation matrices of the time series. This score is robust to the presence of uncertainty; it is not very sensitive to transient changes; it allows capturing complex relationships between time series such as oscillations and trends, and it is also well adapted to the comparison of short time series. The FOTS score is used with the Scalable Unsupervised Shapelet Discovery algorithm for the clustering of 17 datasets, and it has shown a substantial improvement in the quality of the clustering with respect to the Rand Index. This work defines a novel framework for the clustering of uncertain time series.

1 Introduction

Uncertainty in time series comes from several sources. For instance, to protect privacy, privacy-preserving transformation [Papadimitriou *et al.*2007, Aggarwal2008] deliberately introduce uncertainty to the confidential data before further processing. In a sensor network, sensor readings are imprecise because of the presence of noise generated either by the equipment itself or other external influences [Cheng *et al.*2003]. Ignoring the uncertainty of the data can lead to rough or inaccurate conclusions, hence the need to implement uncertain data management techniques.

Several recent studies have focused on the processing of uncertainty in data mining. Two main approaches allow to take uncertainty into account in data mining tasks: either it is taken into account during the comparison by using appropriate distance functions [Rizvandi *et al.*2013, Hwang *et al.*2014, Rehfeld and Kurths2014, Orang and Shiri2014, Wang *et al.*2015, Orang and Shiri2017], or its impact is reduced by transformations performed on the data [Orang and Shiri2015].This latter strategy is used natively by the u-shapelet algorithm.

1.1 U-shapelets algorithm for clustering Uncertain Time Series

U-shapelets clustering is a framework introduced by [Zakaria *et al.*2012] who suggested clustering time series from the local properties of their sub-sequences rather than using their global features of the time series [Zhang *et al.*2016]. In that aim, u-shapelets clustering first seeks a set of sub-sequences characteristic of the different categories of time series and classifies a time series according to the presence or absence of these typical sub-sequences in it.

Clustering time series with u-shapelets has several advantages. Firstly, u-shapelets clustering is defined for datasets in which time series have different lengths, which is not the case for most techniques described in the literature. Indeed, in many cases, the equal length assumption is implied, and the trimming to equal length is done by exploiting expensive human skill [Ulanova *et al.*2015]. Secondly, u-shapelets clustering is much more expressive regarding representational power. Indeed, it allows clustering only time series that can be clustered and do not cluster those that do not belong to any cluster.

Furthermore, it is very appropriate to use u-shapelets clustering with uncertain time series because it can ignore irrelevant data and thus, reduce the adverse effects of the presence of uncertainties in the time series. Despite this advantage, it is highly desirable to take into account the adverse impact of uncertainty during u-shapelet discovery.

1.2 Uncertainty and u-shapelets discovery issue

Traditional measurement of similarity like Euclidean distance (ED) or Dynamic Time Warping (DTW) do not always work well for uncertain time series data. Indeed, they aggregate the uncertainty of each data point of the time series being compared and thus amplify the negative impact of uncertainty. However, ED plays a fundamental role in u-shapelet discovery because it is used to compute the gap, i.e. the distance between the two groups formed by a u-shapelet candidate. The discovery of u-shapelet on uncertain time series could thus lead to the selection of a wrong u-shapelet candidate or to assign a time series to the wrong cluster.

In this study, our goal is to cluster uncertain time series with u-shapelets algorithm. Our work leverages the observation that the use of a dissimilarity function robust to uncertainty could improve the quality of the u-shapelets discovered

and thus improve the clustering quality of uncertain time series.

1.3 Summary of contributions

- We review state of the art on similarity functions for uncertain time series and evaluate them for the comparison of small, uncertain time series.
- We introduce the Frobenius cOrrelation for uncertain Time series uShapelet discovery (FOTS), a new dissimilarity score based on local correlation, which has interesting properties useful for comparison of small, uncertain time series and that makes no assumption on the probability distribution of uncertainty in data.
- We put the source code at the disposal of the scientific community to allow extension of our work [FOTS-SUSh].

2 Definitions and Background

2.1 Related work

An Uncertain Time Series (UTS) $X =< X_1, \ldots, X_n >$ is a sequence of random variables where X_i is the random variable modeling the unknown real value number at timestamp i. There are two main ways to model uncertain time series: multiset-based model and PDF-based model.

In **Multiset-based model**, each element $X_i (1 \leq i \leq n)$ of an UTS $X =< X_1, \ldots, X_n >$ is represented as a set $\{X_{i,1}, \ldots, X_{i,N_i}\}$ of observed values and N_i denotes the number of observed values at timestamp i.

In **PDF-based model**, each element $X_i, (1 \leq i \leq n)$ of UTS $X =< X_1, \ldots, X_n >$ is represented as a random variable $X_i = x_i + X_{e_i}$, where x_i is the exact value that is unknown and X_{e_i} is a random variable representing the error. It is this model that we consider this work.

Several similarity measures have been proposed for uncertain time series. They are grouped into two main categories: Traditional similarity measures and uncertain similarity measures.

- Traditional similarity measures such as Euclidean distance are those conventionally used with time series. They use a single uncertain value at each timestamp as an approximation of the unknown real value.
- Uncertain similarity measures use additional statistical information that quantifies the uncertainty associated with each approximation of the real value : this is the case of DUST, PROUD, MUNICH [Dallachiesa *et al.*2012]. [Orang and Shiri2015] demonstrated that the performances of uncertain similarity measures associated with pre-processing of data are higher than those of traditional similarity measurements.

2.2 Review of u-shapelets

Definition 1 *Two datasets D_A and D_B are said to be **r-balanced** if only if $\frac{1}{r} < \frac{|D_A|}{|D_B|} < (1 - \frac{1}{r}), r > 1$*

Definition 2 *An **Unsupervised-Shapelet** is any sub-sequence that has a length shorter than or equal to the length of the shortest time series in the dataset, and that allows dividing the dataset into two **r-balanced** groups D_A and D_B; where D_A is the group of time series that contains a pattern **similar** to the shapelet and D_B is the group of time series that does not contain the shapelet.*

The similarity between a time series and a shapelet is evaluated using a distance function.

Definition 3 *The sub-sequence distance **sdist(S, T)** between a time series T and a sub-sequence S is the minimum of the distances between the sub-sequence S and all possible sub-sequences of T of length equal to the length of S.*

This definition opens the question of which distance measure to use for $sdist$. In general, the ubiquitous Euclidean distance (ED) is used, but it is not appropriate for uncertain time series [Orang and Shiri2014]. In the following section, we introduce a dissimilarity function that is more adapted to uncertainty.

Computing the $sdist$ between a u-shapelet candidate and all time series in a dataset creates an orderline:

Definition 4 *An orderline is a vector of sub-sequence distances sdist(S, Ti) between a u-shapelet and all time series Ti in the dataset.*

The computation of the orderline is time-consuming. An orderline for a single u-shapelet candidate is $O(NMlog(M))$ where N is the number of time series in the dataset and M is the average length of the time series. The brute force algorithm for U-shapelets discovery requires K such computations, where K is the number of sub-sequences. The strategy used by [Ulanova *et al.*2015] in **Scalable Unsupervised Shapelet algorithm** consists in filtering the K candidate segments by considering only those allowing to build r-balanced groups. This selection is made efficiently thanks to a hash algorithm.

The assessment of a u-shapelet quality is based on its separation power which is calculated as follows :

$$gap = \mu_B - \sigma_B - (\mu_A - \sigma_A), \quad (1)$$

where μ_A (resp. μ_B) denotes mean(sdist(S, D_A)) (resp. mean(sdist(S, D_B))), and σ_A (resp. σ_B) represents standard deviation of $sdist(S, D_A)$ (resp. standard deviation of $sdist(S, D_B)$). If D_A or D_B consists of only one element (or of an insignificant number of elements that cannot represent a separate cluster), the gap score is assigned to zero. This ensures that a high gap scored for a u-shapelet candidate corresponds to a true separation power.

2.3 Review on uncertain similarity functions

Uncertain similarity measures can be grouped into two broad categories : deterministic similarity measurements and probabilistic similarity measurements.

Deterministic Similarity Measures

Like traditional similarity measures, deterministic similarity measures return a real number as the distance between two uncertain time series. **DUST** is an example of deterministic similarity measure.

DUST [Murthy and Sarangi2013] Given two uncertain time series $X =< X_1, \ldots, X_n >$ and $Y =< Y_1, \ldots, Y_n >$, the distance between two uncertain values X_i, Y_i is defined as the distance between their true (unknown) values r(X_i), r(Y_i): $dist(X_i, Y_i) = |r(X_i) - r(Y_i)|$. This distance is used to measures the similarity of two uncertain values.

$\varphi(|X_i - Y_i|)$ is the probability that the real values at timestamp i are equal, given the observed values at that instant :

$$\varphi(|X_i - Y_i|) = Pr(dist(0, |X_i - Y_i|) = 0). \quad (2)$$

This similarity function is then used inside the $dust$ dissimilarity function:

$$dust(X_i, Y_i) = \sqrt{-log(\varphi(|X_i - Y_i|)) + log(\varphi(0))}. \quad (3)$$

The distance between uncertain time series $X =< X_1, \ldots, X_n >$ and $Y =< Y_1, \ldots, Y_n >$ in $DUST$ is then defined as follows:

$$DUST(X, Y) = \sqrt{\sum_{i=1}^{n} dust(X_i, Y_i)^2}. \quad (4)$$

The problem with the deterministic uncertain distances like $DUST$ is that their expression varies as a function of the probability distribution of uncertainty, and the probability distribution of the uncertainty is not always available in time series datasets.

Probabilistic Similarity Measures

Probabilistic similarities measures do not require knowledge of the uncertainty probability distribution. Furthermore, they provide the users with more information about the reliability of the result. There are several probabilistic similarity functions, as MUNICH, PROUD, PROUDS or Local Correlation.

MUNICH [Aßfalg *et al.*2009] This distance function is suitable for uncertain time series represented by the multiset based model. The probability that the distance between two uncertain time series X and Y is less than a threshold ε is equal to the number of distances between X and Y, which are less than ε, over the possible number of distances:

$$Pr(distance(X, Y)) \leq \varepsilon = \frac{|\{d \in dists(X, Y) | d \leq \varepsilon\}|}{|dists(X, Y)|} \quad (5)$$

The computation of this distance function is very time-consuming.

PROUD [Yeh *et al.*2009] Let$X =< X_1, ..., X_n >$ and $Y =< Y_1, ..., Y_n >$ be two UTS each modeled by a sequence of random variables, the PROUD distance between X and Y is $d(X, Y) = \sum_{i=1}^{n} (X_i - Y_i)^2$. According to the central limit theorem [Hoffmann-Jørgensen and Pisier1976], the cumulative distribution of the distances approaches asymptotically a normal distribution:

$$d(X, Y) \propto N(\sum_i E[(X_i - Y_i)^2], \sum_i Var[(X_i - Y_i)^2]) \quad (6)$$

As a consequence of that feature of PROUD distance, the table of the normal centered reduced law can be used to compute the probability that the normalized distance is lower than a threshold:

$$Pr(d(X, Y)_{norm} \leq \epsilon). \quad (7)$$

A major disadvantage of PROUD is its inadequacy for comparing time series of small lengths like u-shapelets. Indeed, the calculation of the probability that the PROUD distance is less than a value is based on the assumption that it follows **asymptotically** a normal distribution. Thus, this probability will be all the more accurate as the compared time series are long (more than 30 data points).

PROUDS [Orang and Shiri2015] is an enhanced version of PROUD, which suppose that random variables coming from time series are independent and identically distributed.

Definition 5 *The normal form of a standard time series $X =< X_1, \ldots, X_n >$ is defined as $\hat{X} =< \hat{X}_1, \ldots, \hat{X}_n >$ in which for each timestamp i $(1 \leq i \leq n)$, we have:*

$$\hat{X}_i = \frac{X_i - \bar{X}}{S_X},\ \bar{X} = \sum_{i=1}^{n} \frac{X_i}{n},\ S_X = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{(n-1)}}. \quad (8)$$

PROUDS defines the distance between two normalized time series $\hat{X} =< \hat{X}_1 ... \hat{X}_n >$ and $\hat{Y} =< \hat{Y}_1 ... \hat{Y}_n >$ (Definition 5) as follows:

$$Eucl(\hat{X}, \hat{Y}) = 2(n-1) + 2 \sum_{i=1}^{n} \hat{X}_i \hat{Y}_i \quad (9)$$

For the same reasons as PROUD, PROUDS is not suitable for short time series comparison. Another disadvantage of PROUDS is that it assumes that the random variables are independent : this hypothesis is strong and particularly inappropriate for short time series like u-shapelets. A more realistic hypothesis with time series would be to consider that the random variables constituting the time series are M-dependent. Random variables of a time series are called M-dependent if $X_i, X_{i+1}, ..., X_{i+M}$ are dependent (correlated) and the variables X_i and X_{i+M+1} are independent. However, the M-dependent assumption could make PROUDS writing more complex and its use more difficult because of the choice of the parameter M.

Uncertain Correlation [Orang and Shiri2017] : Correlation analysis techniques are useful for feature selection in uncertain time series data. Indeed, correlation indicates the degree of dependency of a feature on other features. Using this information, redundant features can be identified. The same strategy can be useful for u-shapelet discovery. Uncertain correlation is defined as follows :

Definition 6 *(Uncertain time series correlation) Given UTS* $X =< X_1, \ldots, X_n >$ *and* $Y =< Y_1, \ldots, Y_n >$*, their correlation is defined as:*

$$Corr(X,Y) = \sum_{i=1}^{n} \hat{X}_i \hat{Y}_i / (n-1), \tag{10}$$

where $\hat{X}_i$ *and* $\hat{Y}_i$ *are normal forms of* X_i *and* Y_i *(Definition 5), respectively.* X_i *and* Y_i *are supposed to be independant continous random variables.*

If we know the probability distribution of random variables, it is possible to determine the probability density function associated with the correlation, which will subsequently be used to calculate the probability that the correlation between two time series is greater than a given threshold. Uncertain correlation has however some drawbacks :

- It is too sensitive to transient changes, often leading to widely fluctuating scores;
- It cannot capture complex relationship in timeseries;
- It requires to know the probability distribution function of the uncertainty or to make some assumption on the independence of the random variables contained in time series.

Because of all thoses drawbacks, uncertain correlation cannot be used as it is for u-shapelet discovery. The next paragraph presents a generalisation of correlation coefficient that is not an uncertain similarity function but is still interesting for u-shapelet discovery.

Local Correlation [Papadimitriou *et al.*2007] is a generalization of the correlation. It computes a time-evolving correlation scores that tracks a local similarity on multivariate time series based on local autocorrelation matrix. The autocorrelation matrix **allows capturing complex relationship** in time series like the key oscillatory (e.g., sinusoidal) as well as aperiodic trends (e.g., increasing or decreasing) that are present. The use of autocorrelation matrices which are computed based on overlapping windows allows **reducing the sensibility to transient changes** in time series.

Definition 7 *(Local autocovariance, sliding window). Given a time series* X*, a sample set of windows with length w, the local autocorrelation matrix estimator* $\hat{\Gamma}_t$ *using a sliding window is defined at time* $t \in \mathbb{N}$ *as (Eq.11) :*

$$\hat{\Gamma}_t(X, w, m) = \sum_{\tau=t-m+1}^{t} x_{\tau,w} \otimes x_{\tau,w}. \tag{11}$$

where $\boldsymbol{x}_{\tau,\omega}$ *is a sub-sequence of the time series of length* w *and started at* τ*,* $x \otimes y = xy^T$ *is the outer product of x and y. The sample set of m windows is centered around time t. We typically fix the number of windows to m = w.*

Given the estimates $\hat{\Gamma}_t(X)$ and $\hat{\Gamma}_t(Y)$ for the two time series, the next step is how to compare them and extract a correlation score. This goal is reached using the spectral decomposition; The eigenvectors of the autocorrelations matrices capture the key aperiodic and oscillatory trends, even **in short time series**. Thus, the subspaces spanned by the first few (k) eigenvectors are used to locally characterize the behavior of each series. Definition 8 formalizes this notion:

Definition 8 *(LoCo score). Given two series* X *and* Y*, their LoCo score is defined by*

$$\ell_t(X,Y) = \frac{1}{2}(\|\boldsymbol{U}_X^T \boldsymbol{u}_Y\| + \|\boldsymbol{U}_Y^T \boldsymbol{u}_X\|) \tag{12}$$

where $\boldsymbol{U}_X$ and $\boldsymbol{U}_Y$ are the k first eigenvector matrices of the local autocorrelation $\hat{\Gamma}_t(X)$ and $\hat{\Gamma}_t(Y)$ respectively, and u_X and u_Y are the corresponding eigenvectors with the largest eigenvalue.

Intuitively, two time series X and Y will be considered as close when the angle α formed by the space carrying the information of the time series X and the vector carrying the information the time series Y is zero. In other words X and Y will be close when the value of the $cos(\alpha)$ will be 1. The only assumption made for the computation of LoCo similarity is that the mean of time series data point is zero. This could be easily achieve with z-normalization. LoCo similarity function has many interesting properties and does not require to:

- Know the probability distribution of the uncertainty,
- Assume the independence of the random variables or the length of u-shapelets.

It is therefore interesting for feature selection, but we still need a dissimilarity function to be able to discover u-shapelet. In the next paragraph, we define a dissimilarity function that has the same properties as LoCo and that is robust to the presence of uncertainty.

3 Our Approach

3.1 Dissimilarity function

The LoCo similarity function defined on two multivariate time series X and Y approximately corresponds to the absolute value of the cosine of the angle formed by the eigenspaces of X and Y ($|cos(\alpha)|$). A straightforward idea would be to use the $sin(\alpha)$ or α-value as a dissimilarity function but this approach does not work so well; the sine and the angle are not discriminant enough for eigenvector comparison for clustering purpose. We thus propose the following dissimilarity measure (Definition. 9).

Definition 9 *(FOTS : Frobenius cOrrelation for uncertain Time series uShapelet discovery) Given two series* X *and* Y*, their FOTS score is defined by*

$$FOTS(X,Y) = \|U_X - U_Y\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{k} (U_X - U_Y)_{ij}^2} \tag{13}$$

where $\|\|_F$ *is the Frobenius norm.*

Because the FOTS computation is based on the comparison of the k-first eigenvectors of the autocorrelation matrices of the time series, it has the same desirable properties of the LoCo similarity function, that is:

- It **allows to capture complex relationship** in time series like the key oscillatory (e.g., sinusoidal) as well as aperiodic (e.g., increasing or decreasing) trends that are present;
- It allows to **reduce the sensibility to transient changes** in time series;
- It is appropriate for the **comparison of short timeseries**.

Moreover, the FOTS dissimilarity function is **robust to the presence of uncertainty** due to the spectral decomposition of the autocorrelation matrices of the time series. The robustness of FOTS to the uncertainty is confirmed by the following theorem:

Theorem 1 ***(HoffmanWielandt)*** *[Bhatia and Bhattacharyya1993] If X and $X + E$ are $n \times n$ symmetric matrices, then :*

$$\sum_{i=1}^{n} (\lambda_i(X + E) - \lambda_i(X))^2 \leq ||E||_F^2. \quad (14)$$

where $\lambda_i(X)$ is the ith largest eigenvalue of X, and $||E||_F^2$ is the squared of the Frobenius norm of E.

The next section explains how FOTS is integrated in the Scalable Unsupervised Shapelet discovery algorithm.

3.2 Scalable u-shapelets Algorithm with FOTS score

In this section we do not define a new SUShapelet algorithm, but we explain how we use SUShapelet algorithm with FOTS score (FOTS-SUSh) to deal with uncertainty.

Two main criteria make possible to evaluate the quality of a u-shapelet:

- It has to produce two r-balanced groups.
- It must build two well separated groups, i.e., groups whose gap is maximal.

The gap is, therefore, an essential criterion for the selection of u-shapelets candidate. It is subject to uncertainty because its calculation is based on the Euclidean distance. To remedy this, we propose to use the FOTS score instead of a simple Euclidean distance when calculating the gap in the Scalable u-shapelet algorithm. Algorithms 1 and 2 present a more formal definition:

Definition 10 *The sub-sequence FOTS dissimilarity $\boldsymbol{sd_f(S, T)}$ between a time series T and a sub-sequence S is the minimum of the FOTS score between the sub-sequence S and all possible sub-sequences of T of length equal to the length of S.*

Algorithm 1: ComputeOrderline

Input: u-shapeletCandidate : s,
time series dataset : D
Output: Distance between the u-shapelet Candidate and all the time series of the dataset

```
function ComputeOrderline(s, D)
    dis ← {}
    s ← zNorm(s)
    forall i ∈ {1, 2, ..., |D|} do
        ts ← D(i, :)
        dis(i) ← sd_f(s, ts)
    return dis/|s|
```

Algorithm 2: ComputeGap

Input: u-shapeletCandidate : s,
timeseries dataset : D,
lb, ub : lower/upper bound of reasonable number of time series in cluster
Output: gap : gap score

```
function ComputeGap(s, D, lb, ub)
    dis ← ComputeOrderline(s, D)
    gap ← 0
    for i ← lb to ub do
        D_A ← dis ≤ dis(i), D_B ← dis > dis(i)
        if lb ≤ |D_A| ≤ ub then
            m_A ← mean(D_A), m_B ← mean(D_B)
            s_A ← std(D_A), s_B ← std(D_B)
            currGap ← m_B − s_B − (m_A + s_A)
            if currGap > gap then
                gap ← currGap
    return gap
```

4 Experimental Evaluation

4.1 Clustering with u-shapelets

The algorithm iteratively splits the data with each discovered u-shapelet: each u-shapelet splits the dataset into two groups D_A and D_B. The time series that belong to D_A are considered as members of the cluster form by the u-shapelet and are then removed from the dataset. A new u-shapelet search continues with the rest of the data until there is no more time series in the dataset or until the algorithm is no more able to find u-shapelet. As a stopping criterion for the number of u-shapelets extracted, the decline of the u-shapelet gap score is examined: the algorithm stops when the gap score of the newly-found u-shapelet becomes less than half of the gap score of the first discovered u- shapelet. This approach is a direct implementation of the u-shapelet definition

Choosing the length N of a uShapelet : The choice of the length of u-shapelet is directed by the knowledge of the domain to which the time series belongs. As part of these experiments, we tested all numbers between 4 and half the length of the time series. We consider as length of u-shapelet the one allowing to better cluster the time series.

Choosing the length w of the windows : The use of overlapping windows for calculating the autocorrelation matrix makes it possible to capture the oscillations present in the time series. During these experiments, we consider that the size of the window is equal to half the length of the u-shapelet.

Choosing the number k of eigenvectors: A practical choice is to fix k to a small value; we use $k = 4$ throughout all experiments. Indeed, key aperiodic trends are captured by one eigenvector, whereas key oscillatory trends manifest themselves in a pair of eigenvectors.

4.2 Evaluation Metric

To appreciate the quality of the u-shapelets found, we use them for a clustering task. The quality of clustering is evaluated from the Rand Index [Rand1971] which is calculated as follows:

Let Lc be the cluster labels returned by a clustering algorithm and Lt be the set of ground truth class labels. Let A be the number of time series that are placed in the same cluster in Lc and Lt, B be the number of time series in different clusters in Lc and Lt, C be the number of time series in the same cluster in Lc but not in Lt and D be the number of time series in different clusters in Lc but in same cluster in Lt. The Rand Index is equals to :

$$Rand\,Index = (A + B)/(A + B + C + D) \qquad (15)$$

4.3 Comparison with u-shapelet

Similarly to [Dallachiesa *et al.*2012], we tested our method on 17 datasets coming from UCR archive [Chen *et al.*2015] representing a wide range of application domains. The training and testing sets have been joined to obtained bigger datasets. Table 1 present detailed information about tested datasets.

Data-set	Size of dataset	Length	No. of Classes	Type
50words	905	270	50	IMAGE
Adiac	781	176	37	IMAGE
Beef	60	470	5	SPECTRO
Car	120	577	4	SENSOR
CBF	930	128	3	SIMULATED
Coffee	56	286	2	SPECTRO
ECG200	200	96	2	ECG
FaceFour	112	350	4	IMAGE
FISH	350	463	7	IMAGE
Gun_Point	200	150	2	MOTION
Lighting2	121	637	2	SENSOR
Lighting7	143	319	7	SENSOR
OliveOil	60	570	4	SPECTRO
OSULeaf	442	427	6	IMAGE
SwedishLeaf	1125	128	15	IMAGE
synthetic_control	600	60	6	SIMULATED
FaceAll	2250	131	14	IMAGE

Table 1: Datasets

Table 2 presents the comparison of the two algorithms.

4.4 Discussion

The use of the FOTS score associated with the SUShapelet algorithm makes it possible to discover different u-shapelets than those found by the Euclidean distance. The FOTS-SUSh improves the results of time series clustering because the FOTS score takes into account the intrinsic properties of the time series when searching for u-shapelets and is robust to the presence of uncertainty. This improvement is particularly significant when the FOTS score is used for the clustering of time series containing several small oscillations. Indeed, these oscillations are not captured by the Euclidean distance but are by the FOTS score whose calculation is based on the autocorrelation matrix. This observation is illustrated by the result obtained on SwedishLeaf dataset.

Datasets	RI_SUSh	RI_FOTS
50words	0.811	**0.877**
Adiac	0.796	**0.905**
Beef	0.897	**0.910**
Car	0.708	**0.723**
CBF	0.578	**0.909**
Coffee	0.782	**0.896**
ECG200	0.717	**0.866**
FaceFour	0.859	**0.910**
FISH	0.775	**0.899**
Gun_Point	0.710	**0.894**
Lighting2	0.794	**0.911**
Lighting7	0.757	**0.910**
OliveOil	0.714	**0.910**
OSULeaf	0.847	**0.905**
SwedishLeaf	0.305	**0.909**
synthetic_control	0.723	**0.899**
FaceAll	0.907	**0.908**

Table 2: Comparison of the Rand Index of SUSH (RI_SUSh) and FOTS-SUSh (RI_FOTS). The best Rand Index is in bold

Time complexity analysis

ED can be computed in $\mathcal{O}(n)$ and FOTS score is computed in $\mathcal{O}(n^\omega)$, $2 \leq \omega \leq 3$ due to the time complexity of the eigenvector decompositions [Pan and Chen1999]. The computation of FOTS score is then more expensive than that of ED. However, its use remains relevant for u-shapelet research as they are often small.

5 Conclusion and Future Work

The purpose of this work was to discover u-shapelets on uncertain time series. To answer this question, we have proposed a dissimilarity score (FOTS) adapted to the comparison of short time series, whose computation is based on the comparison of the eigenvector of the autocorrelation matrices of the time series. This score is robust to the presence of uncertainty, it is not very sensitive to transient changes, and it allows capturing complex relationships between time series such as oscillations and trends. The FOTS score was used with the Scalable Unsupervised Shapelet Discovery algorithm for clustering 17 literature datasets and showed an improvement in the quality of clustering evaluated using the Rand Index. By combining the benefits of the u-shapelets algorithm, which reduces the adverse effects of uncertainty, and the benefits of the FOTS score, which is robust to the presence of uncertainty, this work is defining a framework for clustering uncertain time series. As a perspective to this work, we plan to use the FOTS score for fuzzy clustering of uncertain time series.

References

[Aggarwal, 2008] Charu C Aggarwal. On unifying privacy and uncertain data models. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 386–395. IEEE, 2008.

[Aßfalg *et al.*, 2009] Johannes Aßfalg, Hans-Peter Kriegel, Peer Kröger, and Matthias Renz. Probabilistic similarity search for uncertain time series. In *SSDBM*, pages 435–443. Springer, 2009.

[Bhatia and Bhattacharyya, 1993] Rajendra Bhatia and Tirthankar Bhattacharyya. A generalization of the Hoffman-Wielandt theorem. *Linear Algebra and its Applications*, 179:11–17, jan 1993.

[Chen *et al.*, 2015] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015.

[Cheng *et al.*, 2003] Reynold Cheng, Dmitri V Kalashnikov, and Sunil Prabhakar. Evaluating probabilistic queries over imprecise data. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 551–562. ACM, 2003.

[Dallachiesa *et al.*, 2012] Michele Dallachiesa, Besmira Nushi, Katsiaryna Mirylenka, and Themis Palpanas. Uncertain time-series similarity: return to the basics. *Proceedings of the VLDB Endowment*, 5(11):1662–1673, 2012.

[FOTS-SUSh,] FOTS-SUSh. https://sites.google.com/view/ijcai18fots-sush/.

[Hoffmann-Jørgensen and Pisier, 1976] Jørgen Hoffmann-Jørgensen and G Pisier. The law of large numbers and the central limit theorem in banach spaces. *The Annals of Probability*, pages 587–599, 1976.

[Hwang *et al.*, 2014] Jun Hwang, Yusuke Kozawa, Toshiyuki Amagasa, and Hiroyuki Kitagawa. GPU Acceleration of Similarity Search for Uncertain Time Series. In *2014 17th International Conference on Network-Based Information Systems*, pages 627–632. IEEE, sep 2014.

[Murthy and Sarangi, 2013] Karin Murthy and Smruti Ranjan Sarangi. Generalized notion of similarities between uncertain time series, March 26 2013. US Patent 8,407,221.

[Orang and Shiri, 2014] Mahsa Orang and Nematollaah Shiri. An experimental evaluation of similarity measures for uncertain time series. In *Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS '14*, pages 261–264, New York, New York, USA, 2014. ACM Press.

[Orang and Shiri, 2015] Mahsa Orang and Nematollaah Shiri. Improving performance of similarity measures for uncertain time series using preprocessing techniques. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management - SSDBM '15*, pages 1–12, New York, New York, USA, 2015. ACM Press.

[Orang and Shiri, 2017] Mahsa Orang and Nematollaah Shiri. Correlation analysis techniques for uncertain time series. *Knowledge and Information Systems*, 50(1):79–116, jan 2017.

[Pan and Chen, 1999] Victor Y Pan and Zhao Q Chen. The complexity of the matrix eigenproblem. In *Proceedings of the thirtyfirst annual ACM symposium on Theory of computing*, pages 507–516. ACM, 1999.

[Papadimitriou *et al.*, 2007] Spiros Papadimitriou, Feifei Li, George Kollios, and Philip S Yu. Time series compressibility and privacy. In *Proceedings of the 33rd international conference on Very large data bases*, pages 459–470. VLDB Endowment, 2007.

[Rand, 1971] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[Rehfeld and Kurths, 2014] K. Rehfeld and J. Kurths. Similarity estimators for irregular and age-uncertain time series. *Climate of the Past*, 10(1):107–122, 2014.

[Rizvandi *et al.*, 2013] Nikzad Babaii Rizvandi, Javid Taheri, Reza Moraveji, and Albert Y. Zomaya. A study on using uncertain time series matching algorithms for MapReduce applications. *Concurrency and Computation: Practice and Experience*, 25(12):1699–1718, aug 2013.

[Ulanova *et al.*, 2015] Liudmila Ulanova, Nurjahan Begum, and Eamonn Keogh. Scalable clustering of time series with u-shapelets. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 900–908. SIAM, 2015.

[Wang *et al.*, 2015] Wei Wang, Guohua Liu, and Dingjia Liu. Chebyshev Similarity Match between Uncertain Time Series. *Mathematical Problems in Engineering*, 2015:1–13, 2015.

[Yeh *et al.*, 2009] Mi-Yen Yeh, Kun-Lung Wu, Philip S Yu, and Ming-Syan Chen. Proud: a probabilistic approach to processing similarity queries over uncertain data streams. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 684–695. ACM, 2009.

[Zakaria *et al.*, 2012] Jesin Zakaria, Abdullah Mueen, and Eamonn Keogh. Clustering time series using unsupervised-shapelets. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 785–794. IEEE, 2012.

[Zhang *et al.*, 2016] Qin Zhang, Jia Wu, Hong Yang, Yingjie Tian, and Chengqi Zhang. Unsupervised feature learning from time series. In *IJCAI*, pages 2322–2328, 2016.