CIT. Journal of Computing and Information Technology, Vol. 26, No. 3, September 2018, 179–190
doi: 10.20532/cit.2018.1004217

A Weighted DTW Approach for Similarity Matching over Uncertain Time Series

Liangli Zuo and Li Yan

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China

To measure uncertain time series similarity effectively and efficiently, in this paper, we propose a weighted DTW distance-based approach for uncertain time series with the expected distance. We introduce a weight function to assign weights to a reference point and a testing point. With this function and the WDTW, the accuracy of calculating uncertain time series similarity can be improved. Also, to reduce the storage space and time-consuming, we extend the lower bound function LB_Keogh for DTW into ULB_Keogh for our approach.

ACM CCS (2012) Classification: Mathematics of computing → Probability and statistics → Statistical paradigms → Time series analysis

Information systems → Information retrieval → Retrieval models and ranking → Similarity measures

Information systems → Data management systems → Database design and models → Data model extensions → Uncertainty

Keywords: uncertain time series, similarity matching, dynamic time warping (DTW), weighted DTW

1. Introduction

Time series is an ordered sequence of data and each element of the time series is indexed by a point in time [1]. Time series data widely exist in many application domains such as finance, meteorology [1], biological science [3], [4], astronomy [5], [6] and so on. The commonly obtained time series are daily stock price changes in the stock market, daily temperature readings of a weather forecast, one's heartbeat changes and the audio or image of multimedia that is transformed into time series data.

With the widespread applications of computer technology in various fields, such data is appearing more and more. Therefore, it is of crucial importance to effectively and efficiently manage and use time series data. Time series data processing has attracted much more attention, and much work has been put into providing a more efficient solution to analyze time series data. This is especially true in the fields of scientific and engineering applications. Here we give an example of a recent job. In [5], [6], coronal mass ejection (CME) data are modeled as time series, and the problem of magnetic cloud (MC) or non-MC distinction in CME data is solved by analyzing a time series data in which clustering and visualization of the time series data are investigated. For clustering, the results from the popular hierarchical agglomerative clustering technique to a distance density clustering heuristic in [7] are compared. For visualization, decision trees are applied to aggregate single-dimensional clustering results to form a multidimensional time series decision tree, with averaged time series to present each decision. Similarly, a data-driven approach is proposed in [8] to address the problem of flare prediction from a multivariate time series analysis perspective. The authors in [8] cluster potential flaring active regions by applying Distance Density clustering on individual parameters and further organize the clustering results into a multivariate time series decision tree.

Many problems have been studied in the literature for the analysis of time series data. Among these issues, the similarity matching problem

of time series is the most basic and most important problem. The similarity of time series is not only directly used for similarity search [9] and clustering [10], [8] of time series, but also provides fundamental support for outlier detection, pattern discovery [11], classification [12], segmentation [9] and so on. As a subtask of these technologies, the similarity problem can provide powerful help for time series prediction and analysis.

In time series, it is usually assumed that all values at timestamps are reliable and clear. But this assumption is not always satisfied. In many practical situations, the value at each time point is indeterminate and is described by an indeterminate value. A time series with uncertain data is called an uncertain time series [13]. There are two major reasons for the uncertainty of time series. The first is related to the physical collection device of time series data. For example, the accuracy of data obtained from wireless sensors is associated with a certain error distribution. The second is related to privacy preservation of time series data. For this purpose, a certain degree of uncertainty is sometimes intentionally introduced into a time series. Uncertain time series widely exists in various applications such as data recording of moving objects, weather forecast and sensor network monitoring.

Like classical time series, the similarity matching is also a fundamental and crucial issue in the analysis of uncertain time series data [14], [15]. Some efforts have been made to give solutions to similarity measurements of uncertain time series based on different models of uncertain time series (i.e., [13], [16], [17]). Uncertain time series matching algorithms can be applied in processing data (skyline queries on uncertain time series [18] and in solving some application problems (e.g., CPU utilization time patterns of several MapReduce applications in [19].

Dynamic time warping (*DTW*) [20], [1] and Euclidean distance [9] are two representative distance measures for classical time series. But they do not work for uncertain time series data. To exactly measure similarity of uncertain time series, in this paper, we follow the step of [10] and propose an uncertain weighted *DTW* (*UWDTW*) distance for uncertain time series based on the expected distance. We give a function for weighting between two points and introduce the weight into the distance calculation of the corresponding two points, so that the similar true points is not replaced by the interference data. With the *UWDTW*, the accuracy of calculating the similarity of uncertain time series can be improved. Also, to reduce the storage space and time-consuming, we extend the lower bound function LB_{Keogh} [21] for *DTW* and get the ULB_{Keogh} for the *UWDTW*.

We organize the rest of this paper as follows. In Section 2, we review the related work. We propose a new similarity matching algorithm for uncertain time series in Section 3. In Section 4, we evaluate our approach with experiments. We conclude this paper and give our future work in Section 5.

2. Related Work

2.1. Similarity of Time Series

The Euclidean distance is first used in the time series matching algorithm in [9]. It is easy to understand, simple to calculate, and efficient. The drawback is the lock-step feature, which means that only two points with the same timestamp can be calculated and the two-time series need to be of the same length. Berndt and Clifford first introduced the *DTW* into time series classification in [11]. The *DTW* overcomes the problem that Euclidean distance cannot be matched due to time series distortion. After the introduction of *DTW*, a large number of time series studies were performed based on *DTW*. The results show that the *DTW* has good performance in time series data analysis [11], [12].

Note that the *DTW* does not take into account the effect of time series bending and shifting on distance calculations and this leads to misclassification. This is especially true in the applications where the shape similarity between two sequences is a primary consideration for accurate identification. So, in [22], a novel distance measure called weighted *DTW* (*WDTW*) is proposed. *WDTW* is a penalty-based *DTW*, which can penalize the points with a higher phase difference between two points to prevent minimum distance distortion caused by outliers. Considering the point values and derivatives without extra parameters, Shen *et al*. [23] propose the summation dynamic time series warping (SDTW in [23]). They integrate piece-

wise linear approximation (PLA) and SDTW as PLA-SDTW to reduce the time consumption of SDTW. It exhibits superiority in time complexity, and the similarity measure is demonstrated.

In [24], Roggen *et al*. present and evaluate a microcontroller-optimized limited-memory implementation of a Warping Longest Common Subsequence algorithm (Warping LCSS). The distance refers to the minimum number of edit operations that make conversion between two series, including character substitution, insertion, and deletion. When two-time series have similar patterns in most of the time periods, in other words, two-time series have distortions and breakpoints only in small ranges, LCSS distance measurement can be used.

2.2. Similarity of Uncertain Time Series

Uncertainty in time series observed in daily life is prevalent. There is a growing interest in studying uncertain time series. Distance measurements for uncertain time series are rare [13], [14], [15], [16], [17]. The issue of probabilistic bounded range query (PBRQ) for uncertain time series data is first discussed in [10]. Given a distance bound ε and a probability threshold τ, two uncertain time series are considered to be similar if the probability that the distance between them is less than or equal to ε is greater than or equal to τ. This can be formally described as follows.

$$\begin{aligned} &\mathrm{PBRQ}_{\varepsilon,\tau}\left(T, DB\right) \\ &= \left\{T' \in DB \mid \Pr\left(DIST\left(T, T'\right) \le \varepsilon\right) \ge \tau\right\} \end{aligned} \quad (1)$$

Here DB is a set of uncertain time series, and $DIST$ () means a distance measurement. Note that the approaches in [13], [16] have different definitions for $DIST$ ().

Aßfalg *et al*. [13] use several observations to represent the uncertainty of each point of the time series. Given an uncertain time series T, they calculate the regular time series TS in the way of picking one observation for each element of T. Then the distance measurement is defined as a set of distances between all combinations from two TSs. There are two major problems in their method. First, an uncertain time series T generally corresponds to some regular time series TSs and the time-consuming in the distance calculation is high. Second, not all application domains can provide multiple sample points for each time slot.

Yeh *et al*. use a different data model than [13] and propose an approach named PROUD to deal with the uncertain time series in [16] Instead of using sample points for each element, the elements of each timestamp are treated as random variables with means and deviations. The distance between two points is represented by Euclidean distance. According to the *Central Limit Theorem* [25], the entire distance between two uncertain time series is considered as a normal distribution. Like the approach in [13], the values of τ and ε are provided by a human and affect similarity calculation greatly. But it is generally difficult for humans to give these values.

A new theoretical framework is proposed in [17], which summarizes the concept of similarity between uncertain time series. The proposed algorithm DUST is an approach for computing the exact distance between two uncertain time series. The DUST is mainly used in such cases when different points in the same uncertain time series obey different probability distributions. Although DUST can give an intuitive answer, it degenerates into the Euclidean distance when whole points in one time series hold the same distribution. In this case, the use of a sophisticated distance measure that accommodates uncertainty is not necessary. Given the special conditions, DUST cannot be applied to most time series.

In [13], each point of an uncertain time series is considered as a discrete random variable. In an uncertain time series, one or several certain time series can be extracted to represent the original uncertain time series. Then the distance calculation between uncertain time series turns to the distance calculation between certain time series.

Several major similarity measures for uncertain time series (including the PROUD, MUNICH and DUST), were analyzed and evaluated in [14], [15]. It is shown that they have very different performances depending on the amount of preliminary information (i.e., a priori knowledge of the characteristics of the time series values and the errors of uncertain time

series). Therefore, similarity measures for uncertain time series is still an open issue. Based on the *DTW*, we propose the *UWDTW*, which works for the probabilistic distribution data model. The *UWDTW* is similar to the MINICH constructed for multiset–based uncertain data and it only requires the means and variations at each time point of uncertain time series. This article will show that the *UWDTW* performs better than Euclidean distances.

3. Weighted *DTW* for Uncertain Time Series Similarity

Let $Q = [q_1, q_2, \ldots, q_i, \ldots, q_n]$ and $R = [r_1, r_2, \ldots, r_j, \ldots, r_m]$ be two-time series with length of n and m, respectively. The distance between Q and R is defined as follows [21]:

$$DTW(Q,R) = \min\left\{\sqrt{\sum_{k=1}^{K} w_k}\right\} \quad \left(\max(n,m) \le K \le n+m-1\right) \tag{2}$$

Here w_k represents the k-th element of a warping path W. This warping path can be found by dynamic programming and its dynamic programming recursively evaluates

$$\gamma(i,j) = dist(q_i, c_j) + \min\left\{\gamma(i-1, j-1), \gamma(i, j-1)\right\}$$

Here $dist(q_i, c_j) = (q_i - c_j)^p$ is the distance between two points responding to (i-th, j-th) points of the distance matrix. The best distance measure is related to the selection of p. Even though an optimal p depends on applications, l_1 and l_2 are usually good choices to classify time series dataset [21]. To find the best match between these two sequences, we retrieve a path through the matrix that minimizes the total cumulative distance between them. In particular, the optimal path is the path that minimizes the warping cost.

3.1. Expected Distance

In this section, we propose a new distance measure for uncertain time series. We use a general uncertain time series model in [16], [17], in which the value of uncertain time series at a time point is considered as a random variable. For a random variable at a time point, it contains a mean and a deviation.

An uncertain time series Q_u is a time series that contains uncertain data at time points. Given a time point j, the value of the uncertain time series at the time point j is represented by $Q_u[j]$ and formally defined as $Q_u[j] = q_{uj} + e_{uj}$. Here q_{uj} is the true value and e_{uj} is the error. The error function could be any arbitrary probability distribution. Hence, at each time point j, the element is a random variable with the mean μ_{uj} and deviation σ_{uj}.

The real distributions are generally unknown and the random variables at different timestamps are assumed to be independent. It is showed that the DTW is hard to be defeated [20]. If the DTW is applied to calculate the distance between uncertain time series with the observed values, the distance accuracy cannot be guaranteed with the increasing errors. In this article, we propose a new distance measurement based on DTW, which uses the means and variances of all time points to calculate the distance between uncertain time series. According to [10], we use expected distance to represent uncertain time series distance.

Theorem 1. Let Q and R be two independent uncertain time series. The distance between Q and R is defined as $dist(Q, R) = ED(Q, R)$. Then we have the following expected distance $ED(Q, R)$.

$$ED(Q,R) = \left(E(Q) - E(R)\right)^2 + Var(Q) + Var(R) \tag{3}$$

Here $E()$ and $Var()$ mean the expected value and variance, respectively.

We use the probability calculation method to get the result of $ED(Q, R)$. We get:

$$ED(Q,R) = \iint (q-r)^2 f_{Q,R}(q,r)\,dqdr.$$

The "independent" in the above theorem (Theorem 1) means that the random variables of all timestamps are independent. Therefore, $f_{Q,R}(q, r) = f_Q(q)f(r)$ is satisfied when calculating the distance because the two subjects have unrelated probabilistic density. Then,

$$\iint (q-r)^2 f_{Q,R}(q,r)\,dqdr = \iint (q-r)^2 f_Q(q) f_R(r)\,dqdr.$$

According to the distributive law of multiplication, the equation can be decomposed into three parts for summation. It turns out as:

$$\iint q^2 f_Q(q) f_R(r)\,dqdr + \iint r^2 f_Q(q) f_R(r)\,dqdr - 2\iint qr f_Q(q) f_R(r)\,dqdr.$$

Solving the definite integral of $\iint q^2 f_Q(q) f_R(r)\,dqdr$, because there are no other factors concerning r, the result of $\int f_r(r)\,dr$ is easily obtained, which is 1, then the integral is equal to 1 times $\int q^2 f_q(q)\,dq$, which can be replaced by $E(Q^2)$. Similarly, the answer to the second part is easy to obtain. As for the third part, it can be split into two individual products, each of which meets the definition of the expected value, namely $E(Q)$ and $E(R)$. Finally, we get the polynomial

$$E(Q_2) + E(R_2) - 2E(Q)\,E(R).$$

It is well known that there is a frequently used relationship between $E(Q^2)$ and $(E(Q))^2$, $Var(Q) = E(Q^2) - (E(Q))^2$. So the final answer is:

$$ED(Q,R) = (E(Q) - E(R))^2 + Var(Q) + Var(R).$$

It can be intuitively observed from the above fomula that the expected distance can reflect data uncertainty well. First, $(E(Q) - E(R))^2$ is becoming smaller along with the decreasing distance difference. Second, $Var(Q) + Var(R)$ indicates that the distance between two points is becoming larger along with the increasing errors in the uncertain time series. It is shown in the above proof that the expected distance considers both the mean and variance that are two important parameters of uncertain time series data. It is feasible to describe uncertain time series data with the expectation and variance of uncertain data.

3.2. Weight Function

In the *DTW*, a matrix is used to obtain a cumulative value as the final distance. In this paper, considering the uncertainty of time series, we provide a weight value for each point of this distance matrix. For two uncertain time series, a weight function assigns a weight value to each element on the matrix based on the difference between the two corresponding points. The differences between the points of these two uncertain time series determine their distance. The smaller the difference, the smaller the distance, and vice versa. In uncertain time series, the value at a time point may be uncertain. We argue that in an uncertain time series, the possible values at a time point generally differ from the mean by a small deviation. So, the weight is mostly determined by the difference in the mean.

One of the most popular classical symmetric functions that use only one equation is the logistic function [22]. However, the standard form of the logistic function is not flexible in setting bounds on weights. Therefore, we modify the standard logistic weight function to make it possible to give weights to the alignments of the matrix. We have:

$$weight(i,j) = \frac{w_{\max}}{1 + \exp\left(-g\left(|q_i - r_j|\right)\right)}. \quad (4)$$

Here $w_{\max}$ means the maximum value of weight and g is an empirical constant that controls the slope of the function. The value of g could be from zero to infinity.

In the probabilistic bounded range query [16], [13], the range of distance limit is (0, 1). In the following experiments the data will be disturbed and the deviation is roughly in the range of (0, 2). In the case of a deviation of 2, the maximum value of the distance is 2. So, the value of $|q_i - r_j|$ varies from 0 to 2. We select [0, 0.25, 0.5, 1] as a set of decimal values for g, observe the curve change of the function, then select [1, 3, 5, 7, 9] as a set of integer values of g, and observe the curve change of the function. In Figure 1, we give the values of several g and obtain the corresponding graphics. We determine that the value of g is set to 3 to 5, which can provide appropriate weights to efficiently perform the similarity calculation. The results show that too small g values form a gentle curve, resulting in similar weights for all points, and too large g values form a steeper

curve, resulting in a sharp increase in weights. In the second case, the weights cannot make too much contribution.

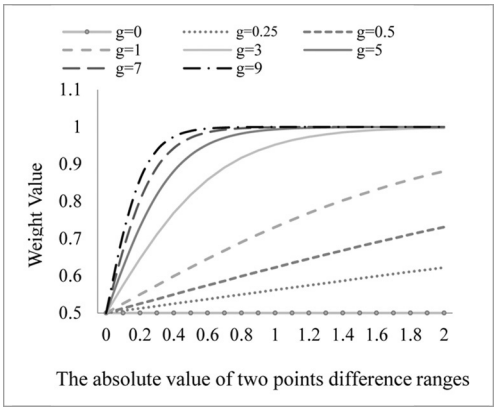

Figure 1. Graphics of weight function with different g ($w_{\max} = 1$).

3.3. Uncertain Weighted *DTW*

Combining the weight function with the expected distance, the optimal distance between two uncertain sequences is defined as follows:

$$UWDTW(Q,R) = dist(q_n, r_m) + \min\{\gamma(n-1,m), \gamma(n,m-1), \gamma(n-1,m-1)\}, \quad (5)$$

$$\gamma(i,j) = dist(q_i, r_j) + \min\{\gamma(i-1,j), \gamma(i,j-1), \gamma(i-1,j-1)\}, \quad (6)$$

$$dist(q_i, r_j) = weight(i,j)\, ED(q_i, r_j). \quad (7)$$

When $i = n$ and $j = m$, the cumulative distance $\gamma(i,j)$ is the dynamic time warping distance that equals to $UWDTW(Q, R)$. As we know, when the deviations of the random variances are large, the use of observed values to calculate uncertain time series distance results in a large error. Here we use the means and variances to represent the random variances and then the calculation results are not greatly affected by the errors.

In the *UWDTW*, two parameters should be pre-assigned before evaluating test performance. Different $W_{\max}$ will not affect its performance. So, we set $W_{\max}$ to 1 in the paper. Also, we choose 3 as the optimal g to give appropriate effective weight. The independent variable in weight function is the absolute difference of means. We use the difference between the mean values to weight each candidate point in the distance matrix. We believe that the average represents the real value. Each point included in the path is intended to minimize the warping cost during the iterative process. The weight has a positive correlation with the independent variable. When the distance between two points is small, the weight will also be small, and the probability that the candidate points are included in the path is greater. Therefore, we can ensure that the candidate points will not be replaced by other points.

3.4. Uncertain LB_{Keogh} (ULB_{Keogh})

Large-scaled time series available can result in high computational cost of processing time series data. To reduce the number of paths to be considered in the calculation process, several well-known constraints are applied to limit the movement that can be made from any point in the path, including *Boundary Conditions*, *Continuity condition*, *Monotonic condition*, and *Adjustment Window Condition* [20]. Two most common global constrains are Sakoe-Chiba Band and Itakura Parallelogram. The Sakoe-Chiba Band uses a pre-set bending radius r value to construct a strip-shaped area as a curved window. The width of this constraint is often set to 10% of the length of the time series. The Itakura Parallelogram uses a fixed algorithm to locate two points in the matrix according to the length of the two-time series and constructs a diamond-shaped area as a curved window, globally limiting the dynamic time curve path.

Based on the warping window, to significantly accelerate the *DTW* calculation, a fast lower bounding technique is used in [20], [21], which can prune sequences that are unlikely to be the best match. For certain time series, there are many lower bound functions of *DTW* distance. Among them, the most famous is LB_{Keogh} proposed by Eammon in [21]. Given two-time se-

ries Q and R with the same length n, the LB_{Keogh} of the two-time series Q and R is defined as follows:

$$LB_{Keogh}(Q,R)=\sqrt{\sum_{i=1}^{n}\begin{cases}(R_i-U_i)^2, if\ R_i>U_i\\(R_i-L_i)^2, if\ R_i>U_i\\0,\ others\end{cases}}. \tag{8}$$

Here $U_i=\max\{Q_{i-r}:Q_{i+r}\}, L_i=\min\{Q_{i-r}:Q_{i+r}\}$ and r is the warp window of DTW.

In the context of uncertain time series, we try to introduce a lower bound function to efficiently perform uncertain time series similarity calculations. We calculate the lower bound function distance of two sequences. If the distance is greater than the minimum distance, this candidate can be omitted and the *WDTW* distance between the two uncertain time series does not need to be calculated. This can reduce the computational cost to some extent.

Based on the expected distance, we propose a novel ULB_{Keogh} function for uncertain time series as follows:

$$ULB_{Keogh}(Q,R)=\sqrt{\sum_{i=1}^{n}\left\{W_i*\left(Var(R_i)+V_i+\begin{cases}(E(R_i)-U_i)^2,\ if\ (E(R_i)>U_i)\\(E(R_i)-L_i)^2,\ if\ (E(R_i)<L_i)\\0,\quad others\end{cases}\right)\right\}}. \tag{9}$$

Here

$$U_i=\max\{E(Q_{i-r}):E(Q_{i+r})\},$$
$$L_i=\min\{E(Q_{i-r}):E(Q_{i+r})\},$$
$$V_i=\min\{Var(Q_{i-r}):Var(Q_{i+r})\},$$
$$W_i=\min\{W(Q_{i-r}):W(Q_{i+r})\},\ W_i=weight(i,i).$$

The expected value of sequence R falls into the closure of U and L. The square of the distance from $E(R)$ to U or L is less than the distance between $E(R)$ and $E(Q)$. The sum of the variance $Var(Q)$ and $Var(R)$ is greater than the sum of the lower bound V and $Var(R)$. Therefore, $ULB_{Keogh}(Q, R)$ is a lower bound function for $UWDTW(Q,R)$.

Theorem 2. Let Q and R be two uncertain time series. For any constrained curved path $j-r\le i\le j+r$, we have

$$ULB_{Keogh}(Q,R)\le UWDTW(Q,R) \tag{10}$$

Proof.

First, we have

$$\sqrt{\sum_{i=1}^{n}\left(W_i*\left(Var(R_i)+V_i+\begin{cases}(E(R_i)-U_i)^2,if\ E(R_i)>U_i\\(E(R_i)-L_i)^2,if\ E(R_i)<L_i\\0,\ others\end{cases}\right)\right)}>\sqrt{\sum_{k=1}^{K}w_k}$$

It is known from the *UWDTW* that $K\ge n$. According to the nature of the warping path, for any i ($1\le i\le n$), there is at least one of the elements of W that is of the form $w_k=(i,j)_k$,

$$w_k=ED(Q_i,R_j)$$
$$=(E(Q_i)-(E(R_j))^2+Var(Q_i)+Var(R_j).$$

For each i, we find a corresponding element w_k in the warping path. For i with multiple matching j, we choose the element with the smallest j. Suppose that these elements constitute a set named A and the rest of elements of w_k constitute a set named B. Then we have:

$$\sum_{i=1}^{n}W_i*\left(Var(R_i)+V_i+\begin{cases}(E(R_i)-U_i)^2,\ if\ E(R_i)>U_i\\(E(R_i)-L_i)^2,\ if\ E(R_i)<L_i\\0,\ others\end{cases}\right)>\sum_{k\in A}w_k+\sum_{k\in B}w_k.$$

Second, assume that $E(R_i) > U_i$, for $j - r \le i \le j + r$, then $i - r \le j \le i + r$, $E(Q_j) \le U_i \le E(R_i)$, $Var(Q_j) \ge V_i$, $W(Q_j) \ge W_i$. Then we have:

$$W_i * \left(Var(R_i) + V_i + \left(E(R_i) - U_i \right)^2 \right)$$
$$\le W(Q_i) * \left(Var(R_i) + Var(Q_j) + \left(E(R_i) - E(Q_i) \right)^2 \right).$$

Similarly, for $E(R_i) < L_i$, we have:

$$W_i * \left(Var(R_i) + V_i + \left(E(R_i) - L_i \right)^2 \right)$$
$$\le W(Q_i) * \left(Var(R_i) + Var(Q_j) + \left(E(R_i) - E(Q_i) \right)^2 \right).$$

For other cases, we have:

$$W_i * \left(Var(R_i) + V_i \right)$$
$$\le W(Q_i) * \left(Var(R_i) + Var(Q_j) + \left(E(R_i) - E(Q_i) \right)^2 \right).$$

So, we have:

$$\sum_{i=1}^{n} W_i * \left(Var(R_i) + V_i + \begin{cases} \left(E(R_i) - U_i \right)^2, if\ E(R_i) > U_i \\ \left(E(R_i) - L_i \right)^2, if\ E(R_i) > L_i \\ 0, others \end{cases} \right)$$
$$\le \sum_{k \in A} w_k$$

and further $\sum_{k \in B} w_k \ge 0$. Finally, we have:

$$\sum_{i=1}^{n} W_i * \left(Var(R_i) + V_i + \begin{cases} \left(E(R_i) - U_i \right)^2, if\ E(R_i) > U_i \\ \left(E(R_i) - L_i \right)^2, if\ E(R_i) > L_i \\ 0, others \end{cases} \right)$$
$$\le \sum_{k \in A} w_k + \sum_{k \in B} w_k.$$

Clearly, it conflicts with the original hypothesis.

The lower bound is used based on global constraints. Being different from the probabilistic approaches proposed in [16], [13], the lower bound is obtained automatically by the function. The thresholds in the probabilistic approaches are provided manually.

4. Experimental Results

In this section, we use the 1NN-classification which is identified as the most appropriate approach for assessing the efficiency of similarity measures [14]. We evaluate our approach on the data sets from "UCR Time Series Data Mining Archive" [26], which is available online. All data sets (including real-life time series, synthetic time series, and generic time series) come from different application domains. These datasets represent time series data, and the time series have been classified into several categories. For each dataset, there is a training set and a testing set. The objective is to perform a 1-NN classification on the testing set to find the nearest time series in the training set. All our experiments are conducted on a PC with a 1.60 GHz CPU and 8 GB of RAM implemented in Java.

Referring to the previous work [2], [4], we artificially add interference to the UCR data to obtain uncertain data. The same processing is performed for all sequences. For the first 10% of the values, we use the normal error function of the standard deviation σ and use the standard deviation of $\sigma/2$ for the next 10%. First, we randomly sample time series, calculate their standard deviations and average them to obtain σ'. Then we multiply each of [0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2] by σ' as the deviation of uncertainty σ_u. Given a σ_u, for each timestamp j, we randomly extract a number from a normal distribution or a uniform distribution as the uncertain value Q_u, where the mean is equal to $Q_u[j]$ and the deviation is equal to σ_u.

We evaluate our approach on 17 datasets and compare the error ratios of our approach to several other approaches, including the traditional *DTW* on original time series (i.e., no error-*DTW*), the *DTW* on perturbed time series, the *DTW* based on *DUST* ($DUST_{DTW}$) and the *DUST* on perturbed data. We compute their similarity results for the maximum standard deviation that is 2. We also analyze and compare the computation time of the *UWDTW* with the *DTW* and *DUST*. We present the results in Table 1, which illustrates the similarity error rates of the five different approaches for each dataset. In this work, the error rate is calculated as follows:

$$Error\ rate = 1 - \frac{total\ number\ of\ correctly\ classified\ data}{total\ number\ of\ testing\ data} \tag{11}$$

For all the other benchmarks, there is a loss rate of close to 10 – 20% in accuracy between the no error-*DTW* and the *DTW*. It can be seen from the table that, for the data sets of Synthetic Control, Face (all), Trace and $Light_7$, our approach has a similar error rate to the $DUST_{DTW}$ and performs better than the $DUST_{DTW}$ on other data sets. Except for the data sets of Synthetic Control, ECG and $Light_7$, for all the other data sets, the *UWDTW* has quite lower error rates than the *DUST*.

In Table 1, the average similarity error ratio is 0.23 for the case of no error *DTW*, 0.24 for the *UWDTW*, 0.29 for the *DUST* and $DUST_{DTW}$, and 0.34 for the *DTW*. We conclude that the *UWDTW* makes more accuracy than the other approaches. In Table 1, we also observe that for the data sets of Coffee and Wafer, the *UWDTW* performs quite well and almost completely makes up for the introduced error.

ULB_{Keogh} is applied in the classification process to reduce the time cost. To classify a time series without a classification label (called unlabeled), we need to make a candidate (called reference) for data items in the dataset being classified. Before calculating the distance between the sequences, we first calculate the lower bound between them and then compare the value to the threshold. The initial threshold is set to be infinity and then gradually replaced with a smaller value. If the lower bound distance is less than the threshold, we continue to calculate the distance between the two uncertain time series, and again compare the distance with the threshold. If it is less than the threshold, we determine that the unlabeled time series and the reference time series have the same label. At the same time, we replace the threshold with this smaller distance value. We select all the elements in the dataset to perform the above process and get the final result. If the lower bound distance is greater than the threshold, the candidate will be directly eliminated without further calculation. In this way, many candidates can be trimmed in

Table 1. Summary of classification performance.

Name	Number of classes	Size of training set	Size of testing set	Time series Length	DTW-No Error (r)	UWDTW	$DUST_{DTW}$	DTW	DUST
Syn_Con	6	300	300	60	0.017 (6)	0.19(3)	0.12	0.2	0.18
GunPoint	2	50	150	150	0.087 (0)	0.160(7)	0.16	0.25	0.18
CBF	3	30	900	128	0.004 (11)	0.07(8)	0.09	0.2	0.2
Face (all)	14	560	1690	131	0.192 (3)	0.32(4)	0.29	0.42	0.35
OSU Leaf	6	200	242	427	0.388 (7)	0.39(7)	0.46	0.48	0.49
Swe_Leaf	15	500	625	128	0.154 (2)	0.3(3)	0.34	0.51	0.3
50Words	50	450	455	270	0.242 (6)	0.30(8)	0.34	0.38	0.39
Trace	4	100	100	275	0.010 (3)	0.18(11)	0.13	0.2	0.22
Two_Patt	4	1000	4000	128	0.002 (4)	**0.05(6)**	0.08	0.31	0.2
Wafer	2	1000	6174	152	0.005 (1)	**0.02(3)**	0.15	0.019	0.02
Light-2	2	60	61	637	0.131 (6)	0.22(12)	0.28	0.22	0.22
Light-7	7	70	73	319	0.288 (5)	0.39(2)	0.37	0.39	0.385
ECG	2	100	100	96	0.120(0)	0.17(4)	0.17	0.22	0.12
Adiac	37	390	391	176	0.391 (3)	0.59(1)	0.78	0.83	0.6
Yoga	2	300	3000	426	0.155 (2)	0.20(5)	0.29	0.32	0.22
Beef	5	30	30	470	0.333 (0)	0.46(2)	0.57	0.5	0.6
Coffee	2	28	28	286	0.000 (0)	**0.10(8)**	0.35	0.42	0.3

advance, and unnecessary distance calculations can be reduced during the calculation. With the ULB_{Keogh} function, the calculation time in the classification process can be well reduced.

We compare the executive time of three approaches: the *UWDTW*, *DTW* and *DUST*. We use the same σ_u as above to randomly extract a number from an exponential distribution to represent the uncertain value. Performing classification of these data and analyzing the time costs, Figure 2 reports the average running time of all data sets for the normal error distribution when the error standard deviation is in the range [0.2, 2.0]. Uniform and exponential distributions have similar results and we do not present these results for brevity. It is shown in Figure 2 that the standard deviation of normal distribution slightly affects performance of the *DUST*. Execution time of the *UWDTW* is not affected at all when the standard deviation for the error of uncertain time series varies. So, the *UWDTW* has the best time performance in all three approaches.

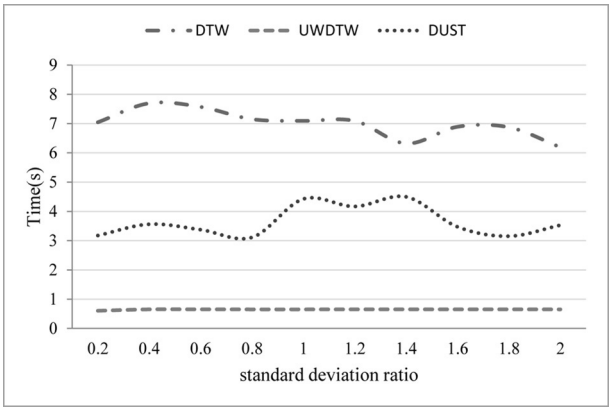

Figure 2. Average execution time of *UWDTW* vs. *DTW* and *DUST*.

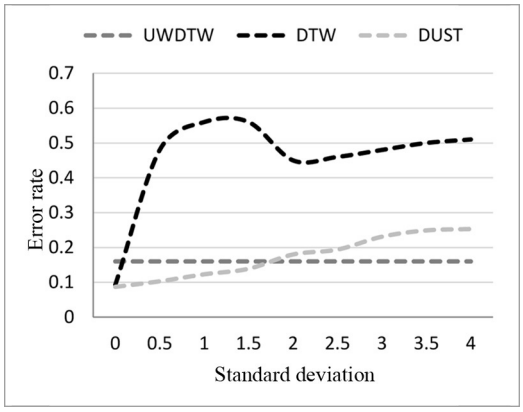

Figure 3. Gunpoint classification error rates with different deviation.

To evaluate the resilient of our approach, we compare the error rates of the *UWDTW*, *DTW* and *DUST* in Figure 3, in which the deviations vary from 0 to 4. It can be seen from Figure 3 that, for the *DUST* and *DTW*, the error rate becomes larger as the error increases. For the *UWDTW*, there is little influence on its ratios. It means that our approach is of a good resilient to the variations.

5. Conclusion

In this paper, we propose a novel *DTW*-based distance approach *UWDTW* to measure uncertain time series similarity. We introduce a weight function to improve the efficiency of this approach. Compared to *DUST* and other probabilistic approaches, *UWDTW* requires less preliminary information and can be easily obtained. The experimental results show that the approach proposed in this paper has better accuracy than the existing uncertain time series similarity methods. Even in the worse cases, when the deviation of uncertain time series is very large, the *UWDTW* can maintain stable accuracy. Also, due to the use of a lower bound function in our approach, the execution time of *UWDTW* does not change much. We evaluate our approach with real data sets and synthetic data sets. As a future work, we will use our approach to classify uncertain time series. We will also explore probabilistic similarity queries on uncertain time series, such as probabilistic nearest neighbor queries.

References

[1] T. M. Rath *et al*., "Word Image Matching using Dynamic Time Warping", in *Proceedings of the Conference of Computer Vision and Pattern Recognition*, 2003, pp. 2–2. http://dx.doi.org/10.1109/CVPR.2003.1211511

[2] L. and J. Xiu, "Application of Satellite Image Time Series and Texture Information in Land Cover Characterization and Burned Area Detection", Doctoral dissertation at University of Helsinki, 2017

[3] M. Lopes *et al*.,"Spectro-temporal Heterogeneity Measures from Dense High Spatial Resolution Satellite Image Time Series: Application to Grassland Species Diversity Estimation", *Remote Sensing*, 2017, pp. 993. http://dx.doi.org/10.3390/rs9100993

[4] D. Calvelo *et al*., "ICU Patient State Characterization using Machine Learning in a Time Series Framework", in *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, Springer, 1999, pp. 356–360. http://dx.doi.org/10.1007/3-540-48720-4_38

[5] R. Z. Ma *et al*., "Solar Flare Prediction using Multivariate Time Series Decision Trees", in *Proceedings of IEEE International Conference on Big Data*, 2017, pp. 2569–2578. http://dx.doi.org/10.1109/BigData.2017.8258216

[6] R. Z. Ma *et al*., "Coronal Mass Ejection Data Clustering and Visualization of Decision Trees", *The Astrophysical Journal Supplement Series*, pp. 4, 2018. http://dx.doi.org/10.3847/1538-4365/aab76f

[7] R. Z. Ma and R. A. Angryk, "Distance and Density Clustering for Time Series Data", in *Proceedings of the IEEE International Conference on Data Mining Workshops*, 2017, pp.25–32. http://dx.doi.org/10.1109/ICDMW.2017.11

[8] R. Z. Ma *et al.*, "A Data-driven Analysis of Interplanetary Coronal Mass Eject and Magnetic Flux Ropes", in *Proceedings of the IEEE International Conference on Big Data*, 2016, pp. 3177–3186. http://dx.doi.org/10.1109/BigData.2016.7840973

[9] C. Faloutsos *et al.*, "Fast Subsequence Matching in Time-series Databases", in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, *ACM*, 1994, pp. 419–429. http://dx.doi.org/10.1145/191843.191925

[10] W. K. Ngai *et al*., "Efficient Clustering of Uncertain Data", in *Proceedings of the International Conference on Data Mining IEEE*, 2006, pp. 436–445. http://dx.doi.org/10.1109/ICDM.2006.63

[11] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series", KDD workshop, 1994, pp. 359–370.

[12] T. Górecki and M. Łuczak, "MultivariateTime Series Classification with Parametric Derivative Dynamic Time Warping", *Expert Systems with Applications*, pp. 2305–2312, 2015.

[13] J. Aßfalg *et al*., "Probabilistic Similarity Search for Uncertain Time Series", in *Proceedings of the International Conference on Scientific and Statistical Database Management*, *Springer*, 2009, pp. 435–443.

[14] M. Orang and N. Shiri, "An Experimental Evaluation of Similarity Measures for Uncertain Time Series", in *Proceedings of the International Database Engineering & Applications Symposium*, *ACM*, 2014, pp. 261–264. http://dx.doi.org/10.1145/2628194.2628207

[15] M. Dallachiesa *et al*., "Uncertain Time-series Similarity: Return to the Basics", in P*roceedings of the VLDB Endowment*, 2012, pp. 1662–1673. http://dx.doi.org/10.14778/2350229.2350278

[16] M. Y. Yeh *et al*., "PROUD: a Probabilistic Approach to Processing Similarity Queries over Uncertain Data Streams", in *Proceedings of the International Conference on Extending Database Technology*, *ACM*, 2009, pp. 684–695. http://dx.doi.org/10.1145/1516360.1516439

[17] S. R. Sarangi and K. Murthy, "DUST: a Generalized Notion of Similarity between Uncertain Time Series", in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 383–392. http://dx.doi.org/10.1145/1835804.1835854

[18] G. He *et al*., "Probabilistic Skyline Queries on Uncertain Time Series", *Neurocomputing*, pp. 224–237, 2016. http://dx.doi.org/10.1016/j.neucom.2015.12.104

[19] N. B. Rizvandi *et al*., "A Study on using Uncertain Time Series Matching Algorithms for MapReduce Applications", *Concurrency and Computation Practice and Experience*, pp. 1699–1718, 2013. http://dx.doi.org/10.1002/cpe.2895

[20] C. A. Ratanamahatana and E. Keogh, "Everything You Know About Dynamic Time Warping is Wrong", Third workshop on mining temporal and sequential data, Citeseer, 2004.

[21] E. Keogh and C. A. Ratanamahatana, "Exact Indexing of Dynamic Time Warping", *Knowledge and Information Systems*, pp. 358–386, 2005. http://dx.doi.org/10.1007/s10115-004-0154-9

[22] Y. S. Jeong *et al*., "Weighted Dynamic Time Warping for Time Series Classification", *Pattern Recognition*, pp. 2231–2240, 2011. http://dx.doi.org/10.1016/j.patcog.2010.09.022

[23] J. Shen *et al*., "A Novel Similarity Measure Approach for Time Series based on PLA and DTW", in *Proceedings of the Chinese Control Conference*, 2016, pp. 7159–7163. http://dx.doi.org/10.1109/ChiCC.2016.7554488

[24] D. Roggen *et al*., "Limited-memory Warping LCSS for Real-time Low-power Pattern Recognition in Wireless Nodes", in *Proceedings of the European Conference on Wireless Sensor Networks*, 2015, pp. 151–167. http://dx.doi.org/10.1007/978-3-319-15582-1_10

[25] E. W. Weisstein, "Central Limit Theorem", 2004. http://mathworld.wolfram.com/CentralLimit Theorem.html

[26] Y. Chen *et al*., "The Ucr Time Series Classification Archive", 2015. www.cs.ucr.edu/~eamonn/time_series_data

Received: December 2017
Revised: June 2018
Accepted: July 2018

Contact addresses:
Liangli Zuo
College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, China
e-mail: liangli_zuo@163.com

Li Yan*
College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, China
e-mail: yanli@nuaa.edu.cn
*Corresponding author

Liangli Zuo is currently a master candidate at the College of Computer Science and Technology at the Nanjing University of Aeronautics and Astronautics, China. Her research interests include time series analysis and uncertain data management.

Li Yan is a full professor at the College of Computer Science and Technology at the Nanjing University of Aeronautics and Astronautics, China. Her current research interests include uncertain data and knowledge engineering.