CIT. Journal of Computing and Information Technology, Vol. 28, No. 4, December 2020, 255–267
doi: 10.20532/cit.2020.1004802

A Hybrid Approach for Clustering Uncertain Time Series

Ruizhe Ma[1], Xiaoping Zhu[2] and Li Yan[2]

[1]University of Massachusetts Lowell, Lowell, MA 01854, USA
[2]Nanjing University of Aeronautics and Astronautics, Nanjing, China

Information uncertainty extensively exists in the real-world applications, and uncertain data process and analysis have been a crucial issue in the area of data and knowledge engineering. In this paper, we concentrate on uncertain time series data clustering, in which the uncertain values at time points are represented by probability density function. We propose a hybrid clustering approach for uncertain time series. Our clustering approach first partitions the uncertain time series data into a set of micro-clusters and then merges the micro-clusters following the idea of hierarchical clustering. We evaluate our approach with experiments. The experimental results show that, compared with the traditional UK-means clustering algorithm, the Adjusted Rand Index (ARI) of our clustering results have an obviously higher accuracy. In addition, the time efficiency of our clustering approach is significantly improved.

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, UK-means clustering, DTW with limited width, hierarchical clustering, ARI

1. Introduction

A time series is an ordered sequence of data and each element of the time series is indexed by a time point. Being one of the most common data types, time series data widely exists in various application fields such as GIS [12], stock market [1], astronomy [21], medical applications [37], meteorology [18], biological science [19]. In addition, some multimedia data (*e.g.*, audio and image data) can be transformed into time series data [32].

Its unique time-dependent and high-dimensional characteristics makes the trend of the property information more visible. As a lot of time series data are becoming available, how to deal with time series data effectively and efficiently is of crucial importance and has attracted more attention. Much work was devoted to proposing various solutions to analyze time series data. For the analysis of time series data, various issues have been investigated in literature. Among these issues, time series data clustering is one of main problems.

Data clustering is one of the essential tasks in data mining, which can provide information on the similar features of objects for analysis. Due to pretreatment steps or subroutines in many other techniques (say, classification [29]), data clustering has gained substantial attention [17]. Currently, many data clustering algorithms have been developed [35], among which the K-means is the most used clustering algorithm that can put the samples into clusters of the nearest cluster center iteratively. In the context of time series data, there are some efforts devoted to time series clustering [4], [13]. In the areas of scientific and engineering applications, for example, time series data clustering especially plays an important role. In [21], 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 clustering and visualizing time series data [20]. Considering the large dimension of given time series, K-means algorithm is used to process time series data due to its high efficiency.

Note that data in practical applications are not always certain and accurate. It is a common case that the initial collected data may contain uncertainty [38]. Several scenarios can result in uncertain data, say accuracy of equipment, location tracking system, personal privacy encryption and so on [38], [30]. Uncertain data widely exist in the real life and uncertain data processing has been hereby investigated in diverse communities. In the context of databases, viewed from the data granularity perspective, we can identify two major categories of uncertain data [38]. The first one is about the uncertainty in objects, which means that it cannot be determined definitely if an object exists, and the second one is about the uncertainty of data values, which means that the attribute values of objects are not accurate. With an increase in the amount of uncertain data available, clustering of uncertain data, which is a crucial issue in uncertain data processing, has become central in uncertain data mining [10]. Some clustering algorithms that tackle these issues have been proposed in the last two decades [3], [17]. Furthermore, there are a few efforts that try to cluster uncertain data streams [2], [15].

Uncertain time series data also exists in many practical applications such as data recording of moving objects, weather forecast and sensor network monitoring. Typically, uncertain time series data can occur in two scenarios. The first one is related to physical collection of time series data. The accuracy of data obtained from a wireless sensor, for example, is associated with a certain error distribution. The second one is related to privacy preservation of time series data and a certain degree of uncertainty is sometimes introduced into a time series intentionally. To deal with uncertain time series, several issues have been investigated, mainly including similarity measurements of uncertain time series based on different uncertain time series models [39], [5], [34], [26], [9] as well as matching and queries [33], [11]. We argue that there are many works on handling uncertain time series as well as clustering crisp data (including common and time series data) and uncertain common data. However, clustering of uncertain time series data is scarcely investigated in the literature. Based on an improved UK-means, in [41], an algorithm of clustering uncertain time series data named "UKMeansULDTW" was proposed. It adopts ULDTW distance with limited width instead of the classical DTW (Dynamic Time Warping) distance to calculate the similarity between uncertain time series and cluster center. But the performance of this clustering algorithm is a problem due to some inherent difficulties in the UK-means clustering [17].

In this paper, we concentrate on clustering uncertain time series data, where the uncertain values at time points are represented by a probability density function (PDF). We calculate the similarity between samples through the probability density. First, with the UK-means clustering algorithm based on the Euclidean distance, uncertain time series data are partitioned into a set of micro-clusters. There is a high degree of clustering in each micro-cluster and the time series within the micro-cluster are very similar. Second, based on the set of micro-clusters, an improved DTW distance is applied as similarity measure to merge the micro-clusters, following the idea of hierarchical clustering, until the number of the target clusters is obtained.

The remainder of this paper is organized as follows. Section 2 presents the related work on time series data clustering, uncertain data clustering and uncertain time series data clustering. In Section 3, we introduce the UK-means clustering algorithm as well as the hierarchical clustering algorithm, and then propose a hybrid approach for clustering uncertain time series by jointly using the UK-means clustering and hierarchical clustering algorithms. Section 4 shows the results of the experiment. Finally, Section 5 concludes this paper.

2. Related Work

Data clustering is an important topic in the research area of data mining. Many efforts have been carried out for clustering time series data and clustering uncertain data.

Considering the large dimension of time series, clustering time series approaches are mainly focused on partitioning and density-based methods. In [6], [22], [23], genetic algorithms were applied for partitional clustering and these approaches have competitive performance in comparison with classical clustering methods. A medoid-based ACO clustering algorithm was proposed by using ant colony optimization in [25], [24]. The K-shape algorithm proposed in [27] is a time series clustering algorithm based on the extensible iterative refinement process. Furthermore, in [13], a hybrid algorithm was proposed based on the advantages of Fuzzy C-means clustering (FCM) and Fuzzy C-medoids clustering (FCMdd). A recent survey on time series clustering is presented in [4].

For clustering uncertain data, the UK-means clustering algorithm was proposed in [7], which is the first uncertain data-clustering algorithm targeting location of devices. The UK-means clustering algorithm applies the probability density function to represent the next possible position of the object. Identifying that the UK-means clustering has some difficulties of time performance and effectiveness because of the uncertainty of objects, some modified UK-means clustering mechanisms were proposed in [17]. In [14], based on probability distribution similarity, an approach for clustering uncertain data was proposed. Along with a plethora of implementations of algorithms, distance measures, indexing techniques, evaluation measures and visualization components, a general framework for clustering uncertain data was proposed in [36]. Instead of a partitional or a density-based clustering approach, the hierarchical clustering paradigm was considered in [10].

Note that most work on uncertain data clustering only considers static uncertain data and ignores the situation that a large number of uncertain data arrives continuously. Few efforts investigate the issue of clustering uncertain data streams (*e.g.*, [2] and [15]). Being very different from static data, time series data is a set of values in order of time, where different time points represent the same attributes and the relative position between them cannot be exchanged. So, the clustering algorithm for static uncertain data cannot be used to cluster uncertain time series directly because the interdependence of time series at time points is not considered and this can result in error clustering results. In addition, time series data is also different from stream data which are a series of data generated continuously. Time points are not the concern of stream data. So, the clustering algorithm for uncertain data streams cannot be directly applied to cluster uncertain time series as well.

Based on the ULDTW distance for similarity calculation, in [41], an improved UK-means clustering algorithm named UKMeansULDTW was proposed for clustering uncertain time series. The UK-means clustering has some inherent difficulties in time performance and effectiveness [17] and, on the other hand, the ULDTW distance has high complexity in similarity calculation. In this paper, we propose an algorithm for clustering uncertain time series data by using the UK-means clustering algorithm and the hierarchical clustering algorithm, which can improve the performance of uncertain time series clustering.

3. Clustering Uncertain Time Series Data with UK-Means and Hierarchical Clustering

The hierarchical clustering includes agglomerative hierarchical clustering and split hierarchical clustering. Being different from the partition clustering that compares the similarity between each sample and the cluster center, the hierarchical clustering focuses on clusters and compares the whole relationships among all clusters during the iteration process. This can weaken the effect of sample monomer on clustering results to some extent. Formally, given a sample set with n data items, say $D = \{x_1, x_2, ..., x_n\}$, let $x_i \in D$ $(1 \le i \le n)$ have m dimensional attributes. Then, by means of agglomeration or splitting, the hierarchical clustering will gradually cluster (partition) D into k sub-clusters, say $C = \{C_1, C_2, ..., C_k\}$, where $D = C_1 \cup C_2 \cup \cdots \cup C_k$. Note that each data item in D can be included in one sub-cluster only. That is, for $x_i \in C_j$ $(1 \le i \le n, 1 \le j \le k)$, we have $x_i \notin C_m$ $(1 \le m \le k$ and $m \ne j)$.

Being different from the partition clustering, it is not needed to set the numbers of final clusters at the beginning of the hierarchical clustering. The numbers of final sample categories are dynamically determined during the hier-

archy partitions according to the indicators of agglomeration or splitting. Figure 1 illustrates the bottom-up processing of agglomerative hierarchical clustering.

In Figure 1, each data item $x_i \in D$ $(1 \le i \le n)$ is initially treated as a separate cluster $C_i^{(0)}$ with a cluster identifier, where (0) means the layer identification of the cluster and i means the order of the cluster at the layer. Then we have $C_1^{(0)}$, $C_2^{(0)}$, ..., $C_n^{(0)}$. We choose two clusters that have the shortest distance (*e.g.*, $C_1^{(0)}$ and $C_2^{(0)}$) and merge them into a new cluster. The other clusters that are not to be merged and the newly generated cluster form an upper layer and are re-assigned new cluster identifiers. Then we have $C_1^{(1)}$, $C_2^{(1)}$, ..., $C_{(n-1)}^{(1)}$, in which $C_2^{(1)}$, ..., $C_{(n-1)}^{(1)}$ correspond to $C_3^{(0)}$, ..., $C_n^{(0)}$, respectively, and $C_1^{(1)}$ is the newly generated cluster formed by merging $C_1^{(0)}$ and $C_2^{(0)}$. We repeat this merging process until a final cluster is obtained or the pre-determined conditions are satisfied. It is shown that the processing of hierarchical clustering forms a tree. Its leaf nodes correspond to the initial statuses of the data items to be clustered, and its root node is the final clustering result. The layer-by-layer changes from leaf nodes to root node reflect the merging process of hierarchical clustering.

In this paper, we cluster the uncertain time series data following the idea of the bottom-up processing of agglomerative hierarchical clustering.. For this purpose, we need to deal with two problems. The first one is how to create the initial clusters for the uncertain time series data. We may treat each original uncertain time series as an initial cluster. But this can result in a large number of initial clusters (it is especially true for massive uncertain time series) and too many iterations, increasing the computing cost. The second one is how to choose two clusters that have the shortest distance for merging from the base clusters. For the first problem, we apply the UK-means clustering algorithm based on the Euclidean distance to create the initial clusters for the uncertain time series data. For the second problem, we introduce an improved DTW distance to calculate the similarity of two clusters.

3.1. UK-Means for Uncertain Time Series

To cluster uncertain data, the traditional K-means algorithm was extended and the UK-means algorithm was proposed in [7]. In the UK-means algorithm, the expected distance is applied to calculate the distance between the samples and the cluster centers. Here, uncertain time series data is a special kind of uncertain data.

Definition. An uncertain time series X is a set of ordered series represented by a tuple $(f(x_i), t_i)$ as follows.

$$X = \{(f(x_1), t_1), (f(x_2), t_2), ..., (f(x_m), t_m)\}, \quad (i = 1, 2, ..., m)$$

Here, $(f(x_i), t_i)$ meats the recorded value at the time point t_i, which is represented as the probability density function. In the practical applications, the probability error function (PEF) is

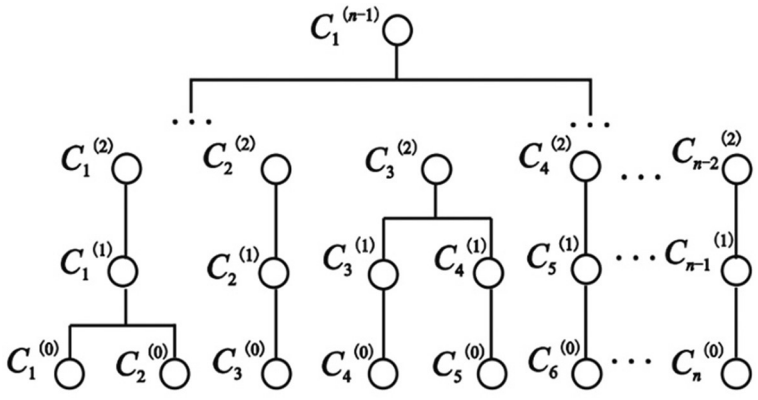


Figure 1. Bottom-up processing of agglomerative hierarchical clustering.

Algorithm 1. UK-means.

Input: uncertain time series set $D = \{x_1, x_2, ..., x_n\}$, the number of micro-clusters K^*, and the max number of iterations T;
Output: the marked sample set C

```
Let numeric (length = n):C                    // record the marker that each sample belongs to
Select K* cluster centers randomly and get c = {c_1, c_2, ..., c_K*}    // initialize cluster centers
t = 0                 // t is used to record the number of iterations
repeat
   for each sample x_i in D do
      minED = Inf
      cIndex = 0
      for each cluster center c_k in c do
         Calculate ED (x_i, c_k)          // calculate the expected distance between x_i and c_k
         if (minED < ED (x_i, c_k)) do
            minED = ED (x_i, c_k)
            cIndex = k
         end if
      end for
      Assign x_i to the cluster represented by c_cIndex and mark C[i] with cIndex
   end for
   Update all cluster centers
   t = t + 1
until no any changes happen in the samples of each cluster or t > T
return C
```

generally applied to represent the difference between the real and recorded values at time points. That is, the real value *real*(*value*) has an error *error*(α) with respect to the recorded value *record*.

Let X be the uncertain time series and c be any cluster center. Then the expected distance between X and c is calculated as follows.

$$ED\left(X, c_j\right) = E\left(\left\|X - c_j\right\|^2\right) = \sum_{i=1}^{m} \int \left\|x_i - c_{ji}\right\|^2 f(x_i) dx_i$$

Here $\| \ \|^2$ represents the Euclidean distance between any two time points and $f(x_i)$ represents the probability density function of X at the time point i. With the formula above, we present the following UK-means algorithm based on the Euclidian expected distance for the uncertain time series data.

By utilizing the efficiency of traditional UK-means clustering algorithm, the sample sets are divided into K^* micro-clusters and then initializing the uncertain time series datasets is completed. When K^* is large enough, the distribution range of each micro-cluster is limited and the samples in each micro-cluster have a similar trend. Then, the distance between the samples and the cluster centers can be directly calculated with the Euclidean distance.

Note that the partition-based UK-means clustering algorithm is simple and highly efficient. But it greatly depends on the initial cluster centers. When the initial cluster centers are not selected properly, the number of iterations possibly increases or the clustering results are only locally optimal. If the cluster centers are randomly selected, we do need to perform clustering several times to verify the clustering algorithm. This procedure increases time consumption. In addition, the agglomerative hierarchical clustering algorithm initially treats each sample as a cluster. As a result, too many iterations might be needed and the calculation time of the algorithm is increased.

3.2. A Hybrid Clustering Approach for Uncertain Time Series

By merging the UK-means clustering algorithm and the agglomerative hierarchical clustering algorithm, we propose a hybrid clustering algorithm named HybridCluster for uncertain time series data. Firstly, we utilize the traditional UK-means algorithm to partition the uncertain time series data into K^* micro-clusters, where K^* is less than the number of actual categories in the samples, so that the samples in each micro-cluster can be clustered as much as possible. Secondly, we utilize the hierarchical algorithm to cluster the K^* micro-clusters instead of the initial samples. This can significantly speed up the iterations of the hierarchical clustering algorithm.

In the hierarchical clustering algorithm, to merge two clusters, we need to calculate the distance between their cluster centers. For this purpose, we introduce an improved DTW distance named ULDTW instead of the original DTW. Compared with the DTW distance, the ULDTW distance can precisely calculate complex similarities of uncertain time series because it can solve the problem of time displacement error between time series [41].

Considering the time complexity of DTW, we apply the method mentioned in [16], which uses a window to limit the width of the matching path in a certain area. Here $path\ [i, j]$ must satisfy the limit that $j - r \le i \le j + r$, where $path\ [i, j]$ records the corresponding points (i, j) of the matching path and r represents the limited width of the window. Finally, we apply the Euclidean distance to the traditional DTW distance and then get the ULDTW distance. The similarity between the uncertain time series and cluster center should satisfy the following:

1. $t = [V][T]$.
2. Ct is temporal class.

Then we have:

$$udtw\left(X_i, Y_j\right) = Ed(i, j) + \min\begin{cases} udtw\left(X_{i-1}, Y_j\right), & i > 1 \text{ and } j = 1 \\ udtw\left(X_{i-1}, Y_{j-1}\right), & i > 1 \text{ and } j > 1 \\ udtw\left(X_i, Y_{j-1}\right), & i = 1 \text{ and } j > 1 \end{cases}$$

Here X_i represents the subsequence $(x_1, x_2, ..., x_i)$ of X. We can see from this formula that in the process of matching the optimal path, the matching path must be incremental in time, which also satisfies the time order of the time series. As a result, the ULDTW distance between the uncertain time series X and the deterministic time series Y can be expressed as follows.

$$ULDTW\left(X, Y\right) = \sqrt{udtw\left(X_p, Y_q\right)}$$

Finally, we have the HybridCluster clustering algorithm described in Algorithm 2, which is more suitable for uncertain time series.

Some discussions of Algorithm 2 are presented as follows.

In the step 2 of Algorithm 2, the selected K^* should be larger than K, the number of actual categories in the sample set, as much as possible. If K^* is too small, on the one hand, the UK-means algorithm can generate some clusters that are too large. The sample distribution within a large cluster is not similar enough and this implies that the samples in the micro-cluster have different categories. On the other hand, too large K^* can result in too many initial clusters for the hierarchical clustering algorithm and this leads to the increase in the number of iterations. In this way, the goal of efficiency improvement by partitioning micro-clusters with the UK-means algorithm cannot be reached. So, we apply $K^* \ge \theta K$ ($K^* \in \mathbb{Z}^+$) to select K^*, where the numbers of sample sets and sample categories are taken into account when θ is selected. Generally, we have $\theta \in$ (1-10). Then, according to the concrete number of categories in the sample set, we can partition the sample into θK micro-clusters so that the sample number in a micro-cluster can be reduced as much as possible and the distribution range of the micro-cluster is as small as possible.

In the step 13 of Algorithm 2, we calculate the cluster centers by using the expected means. Here the value of each cluster center is determinate and the distance between cluster centers can be calculated with the DTW distance.

In the step 19 of Algorithm 2, we adopt the quasi-noise ratio mechanism that uses a single-layer and multi-cluster merge method. With this

Algorithm 2. HybridCluster.

Input: uncertain time series set $D = \{x_1, x_2, ..., x_n\}$, the number of clusters K, the max number of iterations T, convergence threshold θ, and window width r;
Output: the marked sample set C

1. Let numeric (length = n):C // record the marker that each sample belongs to
2. Select proper K^* // initialize the number of micro-clusters
3. C = UK-means (D, K^*, T, θ) // get the set of micro-clusters after partitioning
4. **for** $C_i \in C$ **do**
5. $C_i = C_i^{(0)}$ // initialize the layer number of each micro-cluster
6. Update each cluster center
7. **end for**
8. numOfCluster = K^* // record the cluster number in each layer
9. Let Matrix (ncol = numOfCluster, nrow = numOfCluster):centerDist
   // record the distance between cluster centers
   // numOfCluster is the number of clusters in the current layer
10. **repeat**
11. **for** i in 1: (numOfCluster − 1) **do**
12. **for** j in (i + 1): numOfCluster **do**
13. centerDist [i, j] = DTW (c_i, c_j, r)$dist
    //calculate the distance between any two cluster centers with DTW
14. **end for**
15. **end for**
16. Find from centerDist the sign of the pair of clusters whose centers have a shortest distance
17. (C_i, C_j) = min arg (centerDist)
18. Update the sign of samples in C_i and C_j, and update the center of new cluster
19. Upgrade all clusters to upper layer and their layer number is plus 1
20. Update the DTW distance between center of new cluster and the center of other clusters, and update centerDist
21. numOfCluster = numOfCluster − 1
22. until numOfCluster = K
23. return C

method, a quasi-noise ratio $\overline{\omega}$ is set for cluster merge. The cluster that does not participate in any merge at the current layer is identified as a quasi-noise cluster, which is no longer involved in further merge. Concretely, in the step 19, to upgrade the cluster set at one layer to an upper layer, we identify all cluster centers that have the shortest distances to each cluster and create a set denoted $Min(C) = \{\overline{C}_1, \overline{C}_2, ..., \overline{C}_k\}$. Here $\overline{C}_i$ means the cluster center that has the shortest distance to the cluster center C_i. If there is a cluster center for which the corresponding cluster is not the nearest one among all clusters, it is regarded as the quasi-noise ratio cluster and is no longer involved in the subsequent merge process. This can effectively reduce the sensitivity of the clustering algorithm to noise points and improve the accuracy of clustering.

In the step 20 of Algorithm 2, to update the cluster centers, we calculate the ULDTW distance between each sample and the current cluster center and then calculate new cluster centers according to the matching path *path*. This can detect the time displacement error between the samples in the cluster well. Only the ULDTW distance between the sample in the current cluster and its cluster center is calculated each time, and as a result, less calculations for the ULDTW are needed compared with the UKMeansULDTW algorithm in [41].

4. Experiments

4.1. Construction of Uncertain Time Series

UCR [8] is a time series database which contains many time series sets in different areas, covering a variety of time series sets of different time

dimension, different classes, and different numbers of samples. In order to evaluate the performances of the clustering algorithm proposed in the paper, we utilize the sample sets in UCR as our testing data. For the purpose of making a comparison, we follow the steps described in [14] to construct uncertain time series datasets. We apply different norm distribution functions with expectation of 0 and variance of $0.1\sigma - 0.2\sigma$ as the error function at different time periods, in which σ is the standard variance of each raw sample.

With the construction method of uncertain data above, we can construct uncertain time series as our test datasets. We selected eight sets of data that contain different number of classes as our test datasets, which were also applied in [41]. These test datasets are shown in Table 1. We explain their characteristic parameters: K is the number of sample classes, N is the number of samples of each dataset, and M is the sample dimension (*i.e.*, the time span).

4.2. Experimental Results and Analysis

To compare our algorithm proposed in this paper with the traditional UK-means algorithm and the UKMeansDTW algorithm proposed in [41], Adjusted Rand Index (ARI) [40], which is the same evaluation criterion in [7], is applied to evaluate the clustering results of our clustering algorithm named HybridCluster. We compare each pair of all the samples, and calculate the probability of two samples that belong to the same class and are grouped into the same cluster as well as the probability of two samples that do not belong to the same class and are grouped into different clusters. We set ARI in [−1, 1], where the closer the ARI is to 1, the more accurate the clustering results are.

The experiments in this paper are implemented in R. Table 2 shows the results of the traditional UK-means algorithm, the UKMeansDTW algorithm in [41], and HybridCluster proposed in this paper. It is shown in Table 2 that, compared with the UKMeansULDTW algorithm, the ARI of the HybridCluster algorithm has a little deterioration for partial uncertain time series datasets. But, compared with the traditional UK-means algorithm, the ARI of the HybridCluster algorithm shows significant improvement. For the hierarchical clustering, the HybridCluster algorithm proposed in the paper adopts the ULDTW distance to calculate the distances between samples and cluster centers as well as distances between cluster centers so that the ubiquitous displacement error of time between time series data can be mined. We analyze the reason why the clustering results of the HybridCluster algorithm show a little deterioration for partial uncertain time series datasets with respect to the UKMeansULDTW algorithm. Our findings imply that only the distances between cluster centers and cluster centers are considered when merging clusters with the hierarchical clustering algorithm. At this point, more and more samples in the clusters and their increasingly complex distributions cause the fact that the cluster centers cannot represent all samples fatefully. Thus, errors consequently arise in the cluster merges. The UKMeansULDTW algorithm greatly depends on the initial cluster centers and its clustering results are therefore unstable.

The clustering results of the HybridCluster algorithm, UKMeansDTW algorithm and UKMeans algorithm are evaluated and shown in Table 2. Now we evaluate the time efficiencies of these approaches. Taking into account the uncertain time series datasets that have different sizes and dimensions presented n Table 1, Table 3 shows the comparison of the efficiencies of clustering results by using the UKMeansDTW in [41] and the HybridCluster proposed in this paper.

It is shown in Table 3 that, compared with the UKMeansULDTW algorithm, the time efficiency of the HybridCluster algorithm is significantly improved. In addition, the HybridCluster algorithm obtains the initial clusters of the datasets by using the traditional UK-means algorithm. As a result, the distribution range of initial clusters basically covers the areas of the uncertain time series, and this hereby reduces the sensitivity of the clustering algorithm to randomly selected initial cluster centers. Based on the initial clusters, the samples in similar clusters are gradually merged by using the agglomerative hierarchical clustering algorithm and the time of clustering finally tends to be stable.

Table 1. Experimental datasets.

Dataset	K/N/M	Dataset	K/N/M
Coffe	2/56/286	Plane	7/105/144
BeetleFly	2/40/512	OliveOil	4/30/570
DistalPhalanxOutlineAgeGroup	3/140/80	Symbols	6/25/398
ProximalPhalanxOutlineAgeGroup	3/205/80	Synthetic Control	6/300/60

Table 2. The results of HybridCluster, UKMeansDTW and UK-means.

Dataset	ARI			*K** in HybridCluster	Window width *r* of DTW
	HybridCluster	UKMeansDTW	UKMeans		
Plane 8	0.956	1	0.832	28	5
Coffe 1	1	0.794	0.669	6	3
Symbols 7	0.747	0.753	0.711	8	8
OliveOil 5	0.736	0.718	0.541	8	1
BeetleFly 2	0.734	0.805	0.477	6	1
Synthetic Control 6	0.741	0.77	0.592	15	10
DistalPhalanx-OutlineAgeGroup 3	0.568	0 .611	0 .527	9	5
ProximalPhalanx-OutlineAgeGroup 4	0.606	0 .545	0 .528	18	1

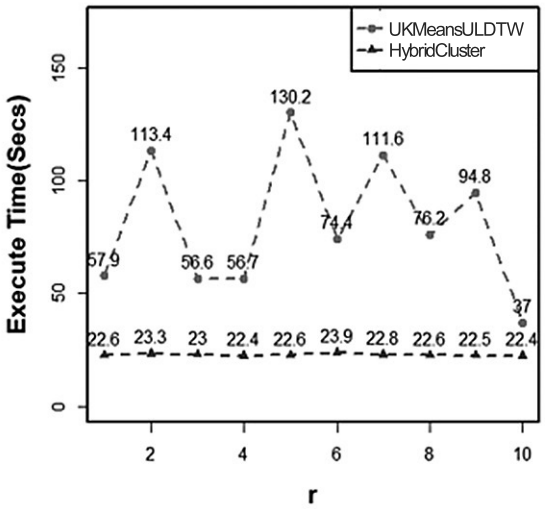
Table 3. The efficiencies of HybridCluster and UKMeansDTW.

Dataset	UKMeansDTW (second)	HybridCluster (second)	*K** in HybridCluster	Window width *r* of DTW	% of improvement
Plane 8	253.8	61.8	28	5	75.7
Coffe 1	74.4	22.6	6	3	69.6
Symbols 7	171.6	37.7	8	8	78.0
OliveOil 5	346.8	50.4	8	1	85.5
BeetleFly 2	193.8	51.4	6	1	73.1
Synthetic Control 6	287.4	32.2	15	10	88.8
DistalPhalanx-OutlineAgeGroup 3	81.7	11.6	9	5	85.8
ProximalPhalanx-OutlineAgeGroup 4	102.6	28.8	18	1	71.9

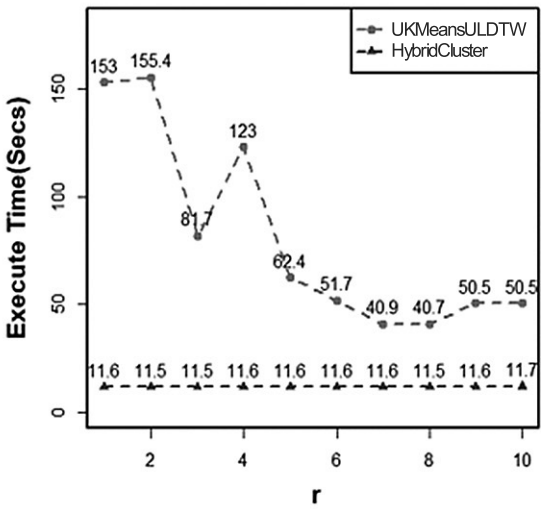
Figure 2 presents the clustering times of using the UKMeansULDTW algorithm and the HybridCluster algorithm after the cluster centers of the same dataset are randomly selected for several times, respectively. It is shown in Figure 2 that when it comes to UKMeansULDTW there is a different cluster center used each time and this results in unsteady iterations, namely, the timing of each execution varies considerably. Furthermore, the execution times of the HybridCluster algorithm are not affected by the randomly selected initial cluster centers and the timing of clustering for the same dataset tends to be stable no matter how many attrempts of random tests are made.

5. Conclusion

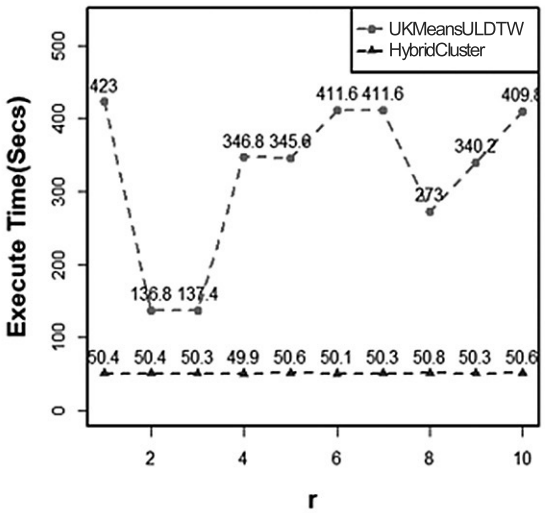
The UK-means algorithm has been widely applied to tackle the problem of clustering uncertain data. In the context of uncertain time series data, uncertain values at time points are represented by probability density functions. For the purpose of clustering uncertain time series data, in this paper, we apply the UK-means



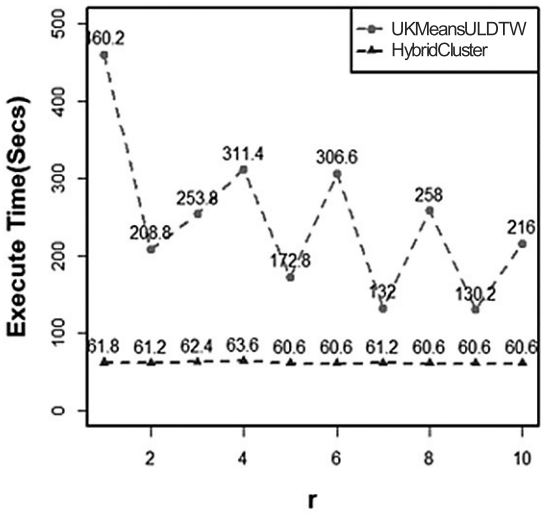
(a) Clustering times for dataset Coffe.



(b) Clustering times for dataset DistalPhalanxOutlineAgeGroup.



(c) Clustering times for dataset OliveOil.



(d) Clustering times for dataset Plane.

Figure 2. Clustering times of the UKMeansULDTW algorithm and the HybridCluster algorithm.

clustering algorithm based on the Euclidean distance to partition the initial uncertain time series data into a set of micro-clusters. Based on the micro-clusters, we apply an improved DTW distance to calculate the similarity of the micro-clusters and then merge the micro-clusters until the number of the target clusters is obtained by following the idea of hierarchical clustering. We demonstrate our clustering approach for uncertain time series data with experiments. The experimental results show that, compared with the traditional UK-means algorithm, the ARI of our clustering approach has a significant improvement. The experimental results also show that, compared with the existing approach for clustering uncertain time series data, the time efficiency of our clustering approach is significantly improved as well. In our future work, we will evaluate and analyze our clustering approach on massive uncertain time series data sets.

References

[1] U. Agarwal and A. S. Sabitha, "Time Series Forecasting of Stock Market Index", in *Proc. of the 2016 India International Conference on Information Processing*, 2016, pp. 1–6. http://dx.doi.org/10.1109/IICIP.2016.7975381

[2] C. C. Aggarwal and P. S. Yu, "A Framework for Clustering Uncertain Data Streams", in *Proc. of the 24th International Conference on Data Engineering*, 2008. pp. 150–159. http://dx.doi.org/10.1109/ICDE.2008.4497423

[3] C. C. Aggarwal, "A Survey of Uncertain Data Clustering Algorithms", Data Clustering: Algorithms and Applications, CRC Press, 2013, pp. 457–482.

[4] S. Aghabozorgi *et al.*, "Time-Series Clustering – A Decade Review", *Information Systems*, vol. 53, pp. 16–38, 2015. http://dx.doi.org/10.1016/j.is.2015.04.007

[5] J. Aßfalg *et al.*, "Probabilistic Similarity Search for Uncertain Time Series", in *Proc. of the 2009 International Conference on Scientific and Statistical Database Management*, Springer, 2009, pp. 435–443. http://dx.doi.org/10.1007/978-3-642-02279-1_31

[6] G. Bello-Orgaz *et al.*, "Adaptive *k*-Means Algorithm for Overlapped Graph Clustering", *International Journal of Neural Systems*, vol. 22, no. 5, 2012. http://dx.doi.org/10.1142/S0129065712500189

[7] M. Chau *et al.*, "Uncertain Data Mining: An Example in Clustering Location Data", in *Proc. of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2006, pp. 199–204. http://dx.doi.org/10.1007/11731139_24

[8] H. A. Dau *et al.*, The UCR Time Series Classification Archive, 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/

[9] M. Dallachiesa *et al.*, "Uncertain Time-Series Similarity: Return to the Basics", *Proceedings of the VLDB Endowment*, 2012, vol. 5, no. 11, pp. 1662–1673. http://dx.doi.org/10.14778/2350229.2350278

[10] F. Gullo *et al.*, "An Information-Theoretic Approach to Hierarchical Clustering of Uncertain Data", *Information Sciences*, 2017, vol. 402, pp. 199–215. http://dx.doi.org/10.1016/j.ins.2017.03.030

[11] G. He *et al.*, "Probabilistic Skyline Queries on Uncertain Time Series", *Neurocomputing*, 2016, vol. 191, pp. 224–237. http://dx.doi.org/10.1016/j.neucom.2015.12.104

[12] H. Izakian and W. Pedrycz, "Anomaly Detection and Characterization in Spatial Time Series Data: A Cluster-Centric Approach", *IEEE Transactions on Fuzzy Syst*, vol. 22, no. 6, pp. 1612–1624, 2014. http://dx.doi.org/10.1109/TFUZZ.2014.2302456

[13] H. Izakian *et al.*, "Fuzzy Clustering of Time Series Data Using Dynamic Time Warping Distance", *Engineering Applications of Artificial Intelligence*, 2015, vol. 39, pp. 235–244. https://doi.org/10.1016/j.engappai.2014.12.015

[14] B. Jiang *et al.*, "Clustering Uncertain Data Based on Probability Distribution Similarity", *IEEE Transactions Knowledge and Data Engineering*, 2013, vol. 25, no. 4, pp. 751–763. http://dx.doi.org/10.1109/TKDE.2011.221

[15] C. Q. Jin *et al.*, "Efficient Clustering of Uncertain Data Streams", *Knowledge and Information Systems*, vol. 40, no. 3, pp. 509–539, 2014 http://dx.doi.org/10.1007%2Fs10115-013-0657-3

[16] E. Keogh *et al.*, "Exact Indexing of Dynamic Time Warping", *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358–386. http://dx.doi.org/10.1007/s10115-004-0154-9

[17] C.-M. Liu *et al.*, "Mechanisms to Improve Clustering Uncertain Data with UKmeans", *Data & Knowledge Engineering*, vol. 116, pp. 61–79, 2018. https://doi.org/10.1016/j.datak.2018.05.004

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

[19] M. Lopes *et al.*, "Spectro-Temporal Heterogeneity Measures from Dense High Spatial Resolu-

tion Satellite Image Time Series: Application to Grassland Species Diversity Estimation", *Remote Sensing*, vol. 9, no. 10, p. 993, 2017. https://doi.org/10.3390/rs9100993

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

[21] R. Z. Ma *et al.*, "Coronal Mass Ejection Data Clustering and Visualization of Decision Trees", *The Astrophysical Journal Supplement Series*, vol. 236, no. 1, pp. 4, 2018. http://dx.doi.org/10.3847/1538-4365/aab76f

[22] H. D. Menéndez *et al.*, "A Genetic Graph-Based Approach for Partitional Clustering", *International Journal of Neural Systems*, vol. 24, no. 3, 2014. http://dx.doi.org/10.1142/S0129065714300083

[23] H. D. Menéndez and D. Camacho, "A Genetic Graph-Based Clustering Algorithm", in *Proc. of the 13th International Conference on Intelligent Data Engineering and Automated Learning*, 2012, pp. 216–225. http://dx.doi.org/10.1007%2F978-3-642-32639-4_27

[24] H. D. Menéndez *et al.*, "MACOC: A Medoid-Based ACO Clustering Algorithm", in *Proc. of the 9th International Conference on Swarm Intelligence*, 2014, pp. 122–133. http://dx.doi.org/10.1007%2F978-3-319-09952-1_11

[25] H. D. Menéndez *et al.*, "Medoid-Based Clustering Using Ant Colony Optimization", *Swarm Intelligence*, vol. 10, no. 2, pp. 123–145, 2016. http://dx.doi.org/10.1007%2Fs11721-016-0122-5

[26] M. Orang and N. Shiri, "An Experimental Evaluation of Similarity Measures for Uncertain Time Series", in *Proc. of the 18th International Database Engineering & Applications Symposium*, 2014, pp. 261–264. http://dx.doi.org/10.1145/2628194.2628207

[27] J. Paparrizos and L. Gravano, "*k*-Shape: Efficient and Accurate Clustering of Time Series", *SIGMOD Record*, vol. 45, no. 1, pp. 69–76, 2016. http://dx.doi.org/10.1145/2949741.2949758

[28] F. Petitjean *et al.*, "A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering", *Pattern Recognition*, vol. 44, no. 3, pp. 678–693, 2011. http://dx.doi.org/10.1016/j.patcog.2010.09.013

[29] G. Pio *et al.*, "Multi-Type Clustering and Classification from Heterogeneous Networks", *Information Sciences*, vol. 425, pp. 107–126, 2018. http://dx.doi.org/10.1016/j.ins.2017.10.021

[30] B. Qin *et al.*, "A Novel Bayesian Classification for Uncertain Data", *Knowledge-Based Systems*, vol. 24, no. 8, pp. 1151–1158, 2011. https://doi.org/10.1016/j.knosys.2011.04.011

[31] J. Qu *et al.*, "Mixed PSO Clustering Algorithm Using Point Symmetry Distance", *Journal of Computer Information Systems*, vol. 20, pp. 53–65, 2010.

[32] T. M. Rath and R. Manmatha, "Word Image Matching Using Dynamic Time Warping", in *Proc. of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003, pp. 521–527. http://dx.doi.org/10.1109/CVPR.2003.1211511

[33] N. B. Rizvandi *et al.*, "A Study on Using Uncertain Time Series Matching Algorithms for MapReduce Applications", *Concurrency & Computation Practice & Experience*, vol. 25, no. 12, pp. 1699–1718, 2013. http://dx.doi.org/10.1002/cpe.2895

[34] S. R. Sarangi and K. Murthy, "DUST: A Generalized Notion of Similarity Between Uncertain Time Series", in *Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 383–392. http://dx.doi.org/10.1145/1835804.1835854

[35] A. Saxena *et al.*, "A Review of Clustering Techniques and Developments", *Neurocomputing*, vol. 267, pp. 664–681, 2017. http://dx.doi.org/10.1016/j.neucom.2017.06.053

[36] E. Schubert *et al.*, "A Framework for Clustering Uncertain Data", *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1976–1979, 2015. http://dx.doi.org/10.14778/2824032.2824115

[37] D. M. Woodbridge *et al.*, "Time Series Discord Detection in Medical Data Using a Parallel Relational Database", in *Proc. of the 2015 IEEE International Conference on Bioinformatics and Biomedicine*, 2015, pp. 1420–1426. http://dx.doi.org/10.1109/BIBM.2015.7359885

[38] L. Xu *et al.*, "Large Margin Clustering on Uncertain Data by Considering Probability Distribution Similarity", *Neurocomputing*, vol. 158, pp. 81–89, 2015. http://dx.doi.org/10.1016/j.neucom.2015.02.002

[39] M. Y. Yeh *et al.*, "PROUD: A Probabilistic Approach to Processing Similarity Queries over Uncertain Data Streams", in *Proc. of the 12th International Conference on Extending Database Technology*, 2009, pp. 684–695. http://dx.doi.org/10.1145/1516360.1516439

[40] S. Zhang *et al.*, "Generalized Adjusted Rand Indices for Cluster Ensembles", *Pattern Recognition*, 2012, vol. 45, no. 6, pp. 2214–2226. http://dx.doi.org/10.1016/j.patcog.2011.11.017

[41] X. P. Zhu *et al.*, "UK-Means Clustering for Uncertain Time Series Based on ULDTW Distance", in *Proc. of the 18th International Conference on Intelligent Data Engineering and Automated Learning*, 2017, pp. 27–35. http://dx.doi.org/10.1007/978-3-319-68935-7_4

Received: May 2019
Revised: April 2020
Accepted: May 2021

Contact addresses:
Ruizhe Ma
University of Massachusetts Lowell
Lowell
USA
e-mail: ruizhe_ma@uml.edu

Xiaoping Zhu
Nanjing University of Aeronautics and Astronautics
Nanjing
China
e-mail: xpzhu@163.com

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

Ruizhe Ma is an assistant professor at the Department of Computer Science at the University of Massachusetts Lowell, USA. Her research interests include deep learning, time series analysis, and Big Data knowledge engineering.

Xiaoping Zhu received her master's degree from 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 management and knowledge engineering.