DTW SIMILARITY MEASURE BASED U-SHAPELETS CLUSTERING ALGORITHM FOR TIME-SERIES DATA

Arathi M [1], Govardhan A [2],

[1]*Research Scholar, Asst. Professor, JNT University Hyderabad, Hyderabad.*

[2]*Professor, JNT University Hyderabad, Hyderabad, Telangana, India.*

([*]**Corresponding author's e-mail:** arathim926@gmail.com)

Document Type:

Submitted: xxx, Accepted: xxx

Abstract

Time-Series Analysis exhibitedefficient results in delivering significant knowledge in numerous domains. Most of the investigationon Time-Series Analysis is restrictedwith the requirementofexpensivecategorized information. This led tothe growth of curiosity in groupingthe time-series informationthat does not need any access to categorized information. The clustering time-series informationcarries out issues that donot prevail in conventional clustering methodologies.,in the Euclidean space amongst the objects.Therefore,the authorsuggested an innovativeclustertechnique, forTime-Seriesemploying of DTW similarity measure by extracting unsupervised shapelets. And these extracted u-shapelets are clustered employing iterative k-means algorithm. The DTW similarity measure provides better accuracy in formed clusters of proposed methodology compared tothe Metric EuclidianDistance Measure. The performance of the suggested approach is evaluated employing theRand Index (RI) Measure. The experimental for this approach was performed on four different Time-Series data samples and the outcomes showed that the RI measure for the DTW based Time-Series Clustering Algorithm is more when compared to the Existing ED-basedTime-Series Clustering Algorithm.

Keywords: Time-Series Analysis, Clustering, Unsupervised Shapelets, K-means, Dynamic Time Warping

1. Introduction

A Time-Series is a gathering of interpretationsrelating to the sequentialarrangement. The environment of time-series information could be categorizedthrough features such as huge data magnitude, highermagnitude, and the necessityforconstant updating. The aim of Time-Series analysis is normally divided into folds: to know or produce the stochastic technique that bounces to aperceivedsequence and to forecast the upcoming values of sequence aidedpastsequence and conceivably other associated series of aspects.The previous decade has perceivedemergingimportance in groupingapproaches for times series data to enableinnovativeappliances in numerous domains like bioinformatics [1,2,3],ecological monitoring [4], financial applications [5], etc.

Clustering Time-Series and other series of informationhave been asignificant domain inspired through numerous studies comprising identical exploration in addition to challenges of producing approaches to distinguish dynamic changes in Time-Series.Times series clustering targetstowardsdetectingconfiguration in an un-categorized Time-Series data samplethroughempiricallyestablishinginformation into homogeneous clusters where the resemblancesamidobjects in a similar group are increased and the resemblancesamidentities across clusters are decreased[6]. The fascinatingshapesconcealed in a Time-Seriessample could be known through reviewing the clusters exposedusing the clustering techniques. Most of the Time-Series clustering approachesdisapprovinglyhinge on the selection of a similarity distance measure, and these procedureshabituallyestimatelikenessesamongstTime-Seriesentities relating tocomplete series. Nevertheless, it is furtherrationale to assess the resemblanceamongsttwoTime-Serieselementspertaining to a fragment of a time intermission alternative to complete series for higher-dimensional Time-Series clustering in numerous appliances.

Time-Series analysis is a vitaldomain inside numerous areas of study comprising aerospace, economics, commerce, weathercasting, medical science, motion capture, etc. [7].Information attained through observations gathered serially over the periodis widely common[8]. Incorporate, weekly interest rates, everydayconcluding stock charges, periodic price catalogs, an annual sales figure, and further are observed. In weathercasting, everyday higher and lowercosts, annual rainfall and faminecatalogs, and hourly wind power are observed. In agriculture, yearlystatistics for harvest and live stocks generation, soil erosion, and export trades are recorded. Inorganic science, the electrical activities at millisecond intermissions are recorded. In an ecosystem, plenty of animal types are recorded. The list of the domain where Time-Series are studied is infinite.

Numerous clustering algorithms are introduced in the earlier few years specifically for Time-Series datasets where the maximum of the clustering techniques employ the linear method for similarity checks such as metric Euclidean Distance, City Block Distance, Manhattan Distance, etc. The approaches give good results for the traditional databases whereas produces very poor results for the Time-Series data samples where the class labels are not known. The linear similarity measures perform better results when the span of different time objects is equal and have no misrepresentation in the time axis. Due to this, the Time-Series data are pruned to obtain equal length for clustering, and the maximum of the evidence is lost.

Therefore, so as to overcome the above-mentioned issue, an innovative cluster technique is introduced by initially extracting the unsupervised shapelets from the unequal Time-Series dataset and employing these u-shapelets for the cluster the Time-Series data. Different from the existing clustering algorithm where linear similarity distances are employed, in this investigation, a non-linear similarity measure is known as Dynamic Time Warping (DTW) is employed to estimate the distance amongst different Time-Series objects. DTW addressed the issue of distortion in the time-axis. DTW techniques evaluate thedistance of two unevensizedTime-Seriesfragments througharranging the signals employing energeticprogram designdepending on constrictions to accommodate those that are identical, however locally out of phase, as shown in Figure 1.

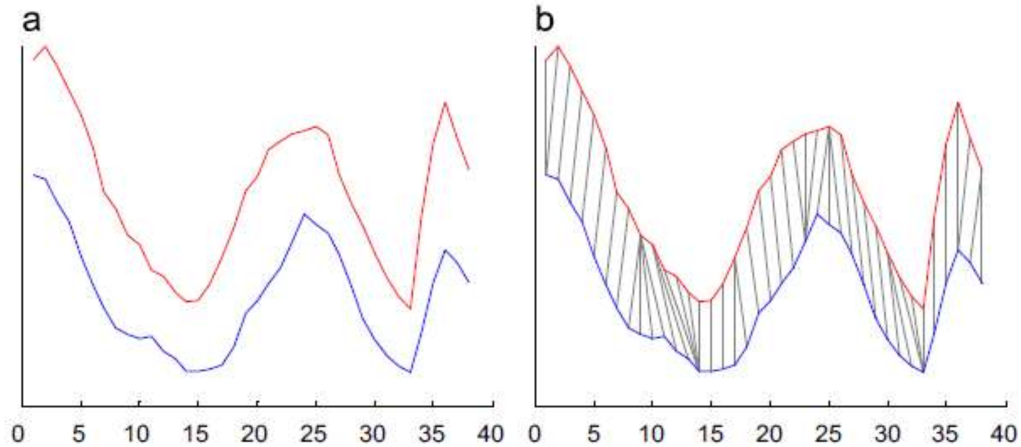

Figure 1: Arrangement of Series using DTW (a) two comparable Sequences out of phase (b) arrangementusing DTW

1.1 Organization of the paper

A brief discussion on the introduction of the Time-Series analysis and motivation for the proposed approach is given in this section. Section 2 gives briefly explains the existing approaches in the Time-Series analysis for clustering using different techniques. The unsupervised shapelets for the cluster of the Time-Series data are briefly explained in section 3. The suggested DTW based U-Shapelets Clustering Algorithm is briefly given in section 4.

2. Literature Survey

Contemporaryapproaches detect shapelets by demanding a pond of applicant sub-series from entireprobable sequential fragments[9, 10] and formerlyarranging the topmost executing fragments pertaining to its destination forecasting abilities. Distances amongst sequence and shapelets exhibit shapelet-transmuted[10] categorization features for a serial fragmentation metrics like information gain [9, 11], FStat [12] or Kruskall-Wallis [13]. The brute-force applicantsexplore methodologies depending on an extensiveexamination of applicants, suffers from a higher execution time complexity, hence numerous rapidly employed approaches are intendedat minimizing the detection period of Shapelets[11, 14, 15]. Identical to stable data clustering, Time-Seriesneeds a clustering procedureto generate groups for the specified group of un-categorized data elements and the selection of clustering approaches hinges on the kind of information accessible and on the specific function and application.

For clustering multimedia Time-Series, [16]accomplished K-means and K-medoid techniques having DTW and showed that K-means is astandard clustering approach whenever Euclidean distance is employed, however,it becomes an unsuccessful attempt to attain accurate outcomes while DTW is employedas a distance measure in summarizing the structure of the Time-Series. As an outcome of the experimentations, this guaranteed that DTW must not be employedlike a subroutine usingthe K-meansapproach and K-medoids usingDTWobtaining adequateoutcomes.

A new technique is defined in [17]for Time-Series clustering where DFT is employed to transfer time-series information from time to frequency channel. Considering, the transmuted amplitude of power band as the characteristics instance of the Time-Seriesinformation, itcould be plotted to a frequency region. OPTICS (Ordering Points to Identify the Cluster Structure) approachis employedto identifygroups in this information.Numerousreplications are specified depending on the past California power marketplace.

In [18], k-Shape, anapproach for time-series grouping is suggested. K-Shape hinges on anascendablerepetitivemodified technique that produced standardized and well-segregated clusters. Considering the distance measure, k-Shape employed a standardizedform of the cross-associationso as to attain the structure of Time-Series whenmatching them. Depending on features of the distance measure, the methodology to estimate cluster centroids are presented that are employed in each generation to upgrade the Time-Seriestasks to groups. Generally, k-Shape is anarea autonomous, precise, and accessiblemethodology.

In [19], a novelvariant of DTW distance measure for Time-Series structure summarizing categorization is suggested. Pertaining to this, resampled DTW and Hybrid DTW provideimprovedaccurateness and higherefficiency when compared to native DTW however to enhance the exactnessadditional, they suggested Contrast-Enhanced DTW (CEDTW) that minimized the influence of data points having non-trivial support to the distance and augmented the efficiencyof the similarity measure.

In [20], a novel procedure for structureaidedTime-Series clustering is presented. It minimizedthe dimension of data, enhanced the effectiveness, and does not diminish the effect through employing the standard of the multifaceted network. Initially, the one-nearest neighbor network is constructed depending on the alikeness amongstTime-Seriesentities. From the adjacent network, every node signifies a single Time-Seriesentity and every association refers to adjoin connection amongst nodes. Subsequently, the nodes having higher degrees are preferred and employed to the group. In the clusteringtechnique, the DTW distance function and hierarchical grouping approach have functioned. Then, certain trails are implemented on artificial and real-world samples. The outcomes exhibited that the suggested technique has worthy performance on competence and efficiency.

Aparsimoniousprototype aided standardforgroupingtimesequenceinformation is given in [28].Inthis technique,theevaluation burdenattains frequently a problempertaining tohugeaccessibleinterpretations.TheestimatedTime-Seriescould likewise belouderandscarceandasuitableprototypeexplainingthemcould becomplextodetermine.The technique is suggested to generatetheviewed measurementsthrough P-splinesmoothersandformerlytocollecttheusefulentitiesprécisedusingoptimumsplinecoefficients.Giving, thefeaturesofwitnesseddimensions,the approach is mergedwiththeappropriategroupingtechnique.The applicationsare offered depending onnon-hierarchicalgroupingapproaches.Theaccuratenessandeffectivenessofthe suggestedareestimatedthroughsimulationsandthroughstudyingtworeal-worlddatainstances.

A novel k -means a form of smooth subspace clustering technique known as Time-Series k -means for clustering is presented in [21]. This algorithm efficiently exploitedintrinsic space data of a Time-Series to improve clustering efficiency. Further precisely, the horizontal subspaces are signifiedwith weighted time stamps thatspecified a comparative discriminative signal of thetimestamps for groupingentities. The foremostassistances of this study comprisethe approach of the novelfitness function to assist the grouping of time-series information and expansion of newapprisingguidelinestogenerategroupexplorationpertaining toeven subspaces. Depending on the artificial samples and five practicalsamples, the experimental outcomesratify that the suggestedprocedureoutstrips other advancedTime-Series clustering approacheswith themutual performance metricslike Accuracy, Fscore, RandIndex, and Usual Mutual Information.

3. Time-Series Unsupervised Shapelets (U-Shapelets)

Time-Series Shapelets, a data mining primitive are Time-Seriessub-sequences which in a certainmannerexcellentlycharacteristic of a label[9] and discriminativesubsequences of Time-Series that finelyforecast the destinationvalue. Shapelets could offer interpretable outcomes that might assist the region expert's better knowledgeof its information. It could be considerablyfurther accurate and strong on certain data samples. This is since they are local characteristics,whilemaximum state-of-art Time-Series or shape classifiers/clusteringfunctions onuniversalcharacteristics that could be delicate to even lower levels of noise and misrepresentations[22].Furthermore, detecting shapelets is aburning area in the time-series domain for the period of the previous five years [10, 11, 14].

The shapelets couldlikewise be extremelyeconomical in clustering the time-series information. Sinceclasses of Time-Series for clustering samples are not known, this leads to the issue of how the shapelets that can be discovered from data samples deprived of having informationon the class labels. This issue proposed a novelkind of shapelet known as unsupervised-shapelet (or u-shapelet) and demonstrated theeffectivenessofgrouping time-series information[23]. Since labels of Time-Series cannot be used to identify the shapelets, a description of informative subsequences is known as unsupervised-shapelets to distinguish them from the traditional shapelets that presumeadmittance to class names. The innovation of u-shapelets is its knowingly disregardingmaximum information, and merely employ a small number of subsequences for grouping. The advantages of U-Shapelets over shapelets are:

1. It is determined for data samples where the discreteelements are of diverse lengths.
2. It alleviates sensitivity to inappropriateinformation like noise, spikes, dropouts, etc.
3. This revealedthe capability to offeradded insights into the data.

3.1 Definition and Background

Definition 1: "*Time-Series*, $TS = TS_1, TS_2, TS_3, \dots . TS_n$is awell-arrangedcollection of physical values. The real values areequivalent to the dimension of Time-Series. A data sample$D = \{TS_1, TS_2, TS_3, \dots . TS_n\}$is a group of N Time-Series. From [23], it is inferred that the small subsequences that best represents its clusters might giveimprovedbundling results compared to employing whole Time-Series.

Definition 2: *Subsequence*, a subsequence$SQ_{i,l}$, where $1 < l < n$and$1 < i < n$, is a bunch of $l$uninterrupted real values from Time-SeriesTS and initiates at location$i$.Time-Series of size$n$has $\frac{n(n+1)}{2}$subsequences of entire probable dimensions. For NTime-Series in the data sample having the size n, formerly there are $N \times \frac{n(n+1)}{2}$ subsequences.

Definition 3: The *Subsequence distance*amongst a subsequence SQ of size$m$ and a Time-SeriesTS of size$n$ is the distance amongstSQ and the subsequence of TShaving minimal distance. It is referred to as $sqdis(SQ, TS)$.

$$sqdis(SQ, TS) = min_{1 \le i \le n-m} dis\left(SQ, TS_{i,m}\right); \quad [1 \le m \le n] \qquad (1)$$

Definition 4: An *unsupervised-shapelet* is a subsequence of a Time-Series$TS$where the sdists amongst Time-Series from a group $D_x$are reduced compared to the sdists amongst and remaining groupD_y in the data sample D.

$$sqdis(\grave{S}Q, D_x) \ll sqdis(\grave{S}Q, D_x) \qquad (2)$$

The matrix comprises of the $sqdis$ vectors amongst u-shapelets and every Time-Series in data sample at a Distance map.

Definition 5: A *Distance map*includes $sqdis$ amongst every u-shapelets and entireTime-Series in the data sample. If there are$m$ u-shapelets for a data sample of NTime-Series, the dimension of the distance map is $[N \times m]$ where every column is a distance-vector of a u-Shapelets"[24].

4. DTW Based U-Shapelets Clustering Algorithm

In the paper, the author presented a novel clustering technique specifically for various kinds of Time-Series data samples. Different from existing shapelets that are employed for the classification of Time-Series data samples, in this approach, unsupervised Shapelets(u-Shapelets) [23]are used merely for the clustering of the data samples. The unsupervised shapelets are extracted without the useof any class labels for the data samples. In this approach, every entity in the samples is a Time-Series, which are of dissimilar size. Initially, as to perform the clustering technique, u-shapelets are extracted from the numerous time-series data samples. This technique clusters the data based on the u-shapelets extracted from the entire Time-Series datasets using the DTW similarity measure and explicitly allows us to ignore some of the information that is not extracted by the u-shapelets such as outliers.The proposed approach is represented in the Figure.2.The proposed clustering algorithm is comprised of two different phases. They are

1. Extraction of U-Shapelets by ignoring the outliers

2. Clustering of Time-Series data using DTW Similarity Measure

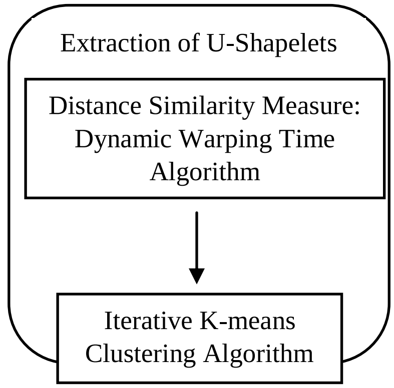

Figure 2: Time-Series Clustering using U-Shapelets with DTW measure

The novelty of this clustering algorithm lies in both the extraction of u-shapelets and the clustering of data. For both these phases, it is essential to use an efficient similarity measure. As given in the previous section that, it is beneficial to employ u-shapelets for clustering purposes instead of simply using the different similarity measures. A similarity measure on Time-Series data series hereafter is further complex to determine compared to traditional informationsince the order of objects in the seriesneeds to be considered. Thus, a non-linear similarity measure is used in the proposed approach instead of linear Euclidian Distance.

4.1 Computation of DTW similarity Measure

The DTW technique depends on the Levenshtein distance and was suggested for appliances in speech recognition. It defines the optimumarrangementamongst two series of arithmeticalvalues and seizesstretchable resemblancesthrough aligning the coordinates within both series. However, DTW does not fulfill the Triangle Inequality property of a Metric Distance Measure, it is useful in comparing time-shifted, stretched, or dilated two unequal length signals. Considertwo-timesequences P and Q of size M and N correspondingly. Let $P = P_1, P_2, P_3, \dots. P_M$ be a series of M examples of the 1-dimensional indicator. Let $Q = Q_1, Q_2, Q_3, \dots. Q_N$be a series of N examples of the other 1-dimensional indicator. DTW uses 2 stages to calculate the warppathway between the 2 sequences. Stage 1 consequences in a matrix of local costs$C(i, j)$ of magnitude$NxM$by means of Minkowski distance measures as in Equation 1. Stage 2 consequences a matrix of universalcosts$D(i, j)$ of magnitudeNxM using local distances as in Equations 2,3, and 4. In stage 1, the Local cost matrix is estimated employing Minkowski distance measure as " :

$$C(i, j) = \|P_i - Q_j\| \qquad (1)$$

In stage 2, the initialrow of the universal distance matrix is evaluated through accumulative summing of local distances of the row as:

$$D(1, j) = \sum_{k=1}^{j} C(P_1, Q_k) \quad for\ j \in [1, M) \qquad (2)$$

Then the First column of the global distance matrix is calculated bythe cumulative sum of local distances of the column as in Eq.3

$D(i,1) = \sum_{k=1}^{j} C(P_k, Q_1) \quad for\ i \in [1, N)$ (3)

For the remaining elements, for i in $[2, N]$, and j in $[2, M]$, the global distance matrix is calculated as in Eq. 4.

$D(i,j) = \min[D(i=1, j-1), D(i-1, j), D(i, j-1)] + c(P_i, Q_j)$ (4)

A warping path W is a contiguous set of matrix elementsthat defines a mapping between two-Time-Series' P and Q.After computing the Global distance matrix, a warped path isfound from $D(1,1)$ to $D(N,M)$"[25].

For any data sample, once the distance map of the u-shapelets' distances is obtained, it could flexibility move into astandard clustering approach known as k-means. Therefore, the concentration ofthestudyis in approachestoattaining highervalued distance maps. Consider that the distance map demonstration is slightlyidentical to the vector space prototype, which is a keystonefor maximum text mining approaches[26].

4.2 Extraction of U shapelets

The higher-level notion of the proposed approach is that it searches for the u-shapelets that could divide and eradicate a subset of Time-Series from remaining data samples, the generativelyrecaps this explorationamongst the remaining information till no data leftoversas separated. The algorithm for the extraction of u-shapelets is given where the distance between the subsequences is evaluated using DTW similarity Measure. This algorithm is viewed as a greedy search approach that tries to exploit the departure gap amongst two subsets of D. Algorithm 1 gives the approach for the extraction of u-shapelets from the Time-Series data sets and Algorithm 2 specifies the approach for the computation of distances between the Time-Series objects using u-shapelets employing DTW algorithm.

The maximum gap is computed and dt of a candidate u-shapelets where the distance vector of applicant u-shapelets. Then, the gap is computed for every possible location dt. The distance vector consumes a series and data sample as input and computes sequence distance amongst subsequence and everyTime-Series in the data sample.

Algorithm 1: Extraction of U-Shapelets

Input: D: dataset, stlen: Shapelet Length

1. Initialize the set of u-shapelets as an empty sequence$\grave{SQ} = []$
2. A Time-Series(ts) dataset is considered $ts = D(1, ;)$
3. While true
    a. $count = 0$
    b. $\grave{sq} = []$
    c. For $seql = sqlen(1)\ to\ sqlen(end)$

i. For $i = 1\ to\ |ts| - seql + 1$

$\grave{sq}(count + 1) = ts(i: i + sl - 1)$

$[sgap(count + 1), dt(count + 1) = (computesGap(\acute{sq}(count + 1), D)$

d. $ind = \max(sgap)$

e. $\grave{SQ} = \grave{sq}(ind)$

f. $distance = compute\ distance(\grave{sq}(ind), D)$

g. $D_A = compute\ (distancne < dt)$

h. If $len(D_A) == 1\ break$

i. Else

i. $index = \max(distance)\ , ts = D(index)$

ii. $\emptyset = mean(distanceD_A) + std(distanceD_A)$

4. Return $\grave{SQ}$

Algorithm 2: Computing the distance between shapelets and dataset

Input: $\grave{sq} = subsequence, D = Dataset$

Output: $distance = distances\ from\ the\ whole\ sequence\ in\ the\ data\ sample.$

1. $distance = []$
2. $\grave{sq} = zNorm(\grave{sq})$
3. $for\ i = 1\ to\ |D|$

a. $ts = D(i,:); distance(i) - inf$

b. $for\ j = 1\ to\ |ts| - |\grave{sq}| + 1$

i. $z = zNorm(ts_{j,|\grave{sq}|})$

ii. $d = DTWDistance(z, \grave{sq})$

iii. $distance(i) = \min\big(d, distance(i)\big)$

4. Return $distance/sqrt(|\grave{sq}|)$

4.3 Clustering of Time-Series

The clustering of the Time-Series dataset with the extracted unsupervised shapelets is done employing the iterative k-means approach by computing objective function employing DTW and keeping clusters that minimize it. The algorithm for clustering of Time-Series is given algorithm 3.

Algorithm 3: Clustering the dataset using iterative k-means approach

Input: D: Data sample,$\grave{SQ} = group\ of\ u - shapelets$, k=number of clusters

Output: $cluster$: $Cluster\ label\ for\ every\ time\ seies$

1. $Dis_{MAP} = []$
2. $Cluster(0) = c$
3. For $count = 1\ to\ length(\grave{SQ})$
    a. $\grave{sq} = \grave{SQ}$
    b. $distance = [Dis_{MAP}\ distance], sumDis_{MAP} = inf$
    c. For $i = 1\ to\ n$
    d. $[Ind, SumD] = k - means(Dis_{MAP}, k)$
    e. $if\ sum(SumD) < sumDis$
        i. $sumDis = sum(SumD)$
        ii. $cluster(count) = Ind$
    f. $RI_c = 1 - RI\big(Cluster(count-1), Cluster(count)\big)$
4. $a = \min(RI_c)$
5. Return $Cluster(a)$

5. Experimental Results and Its Analysis

The experimental results for suggestedTime-Seriesclustering using u-supervised shapelets are performed on the using Matlab r14b version. The time-series data sets that are considered in this suggested technique are Four Classes, Trace, PAMAP, and Birds that are collected from different data repositories. For further details and data, itself is also found in supporting page [27].The complete information pertaining to the experimented data samples are given in Table 1.

Table 1: Experimental Data Samples

Dataset	No. of Time-Series	Time-Series Length	No. of Classes
Four Classes	200	275	2
Trace	200	275	4
PAMAP	345	500	7
Birds	177	500	2

Numerous Diverse metrics for estimating the eminence of Time-SeriesgroupingcomprisesJaccard Score, F1-Score, Rand Index, Folkes and Mallow Index, etc. In this proposed approach, the Rand Index is considered for analyzing the proposed approach. This Index is the most generallyemployed clustering quality metric.

Consider $C = C_1, C_2, \ldots. C_m$is a grouping outline of sample $X$and $P = \{P_1, P_2, \ldots. P_s\}$is a determinedsegment. Collection of points(X_v, X_u) are referred from the sample employing the below terms:

- SS: if both points fita similargroup of the clustering outline C and to a similar set of fragments P.
- SD: if points fitsimilarto group C and diverse set P.
- DS: if points fitdiversegroup C and to similar set P.
- DD: if both points fit diversegroup C and to diverse groups P.

Let us assume that a, b, c,and d represent the values for SS, SD, DS, and DD correspondingly. Here a+b+c+d=M which is the maximal of the entiredata sample. Thus, the Rand Index (RI) or Rand Statistics that measure the degree of similarity amongst C and P is computed as:

$$RI = {}^{(a+d)}\!/_{M}$$

The outcome given does not specify the execution time for the algorithms, since the approach optimizes only for the Quality not for time. The underlying clustering algorithm used is k-means where the accuratelyright value of k is given and reported the finest of twenty executions. The finest gives the execution that diminished the objective function, prior to the classes are seen and Rand Index is computed. Table2 showed the clustering quality of u-shapelets employing existing Euclidean distance measure and proposed DTW distance measure. From the table, it can be inferred that the rand index quality measure for the proposed DTW based U-Shapelets Time-Series clustering is high when compared to the existing ED-based U-Shapelets Time-Series clustering. The Rand Index measure equal to 1 signifies a good clustering approach in the Time-Series analysis. The best U-Shapelets extracted from the proposed approach for the Four Classes, Traces, and Birds Data Samples are shown in Figure 3, Figure 4, and Figure 5.

Table 2: Comparison of Rand Index Values

Dataset	Rand Index	
	U-Shapelets with Euclidean Distance	U-Shapelets with DWT measure
Four Classes	1	1
Trace	1	1
PAMAP	0	
Birds	0.642	0.752

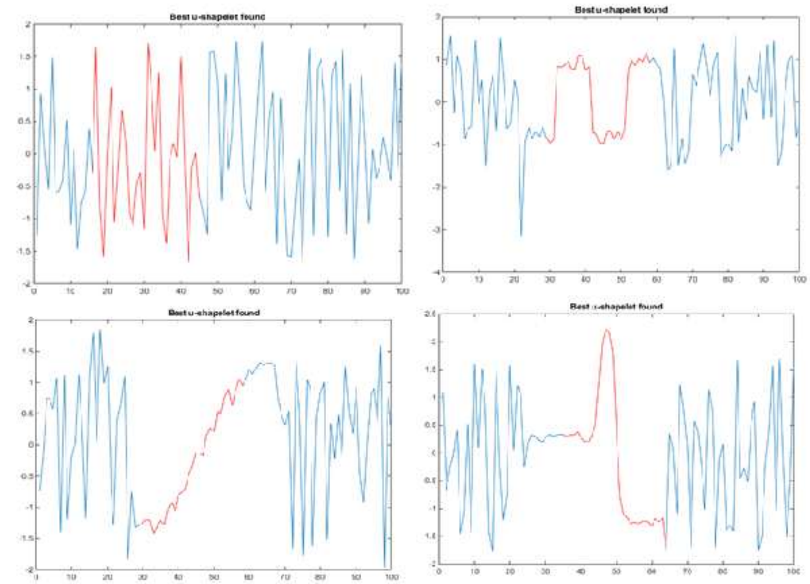

Figure 3: Best U-shapelets found in Four Classes Data Sample

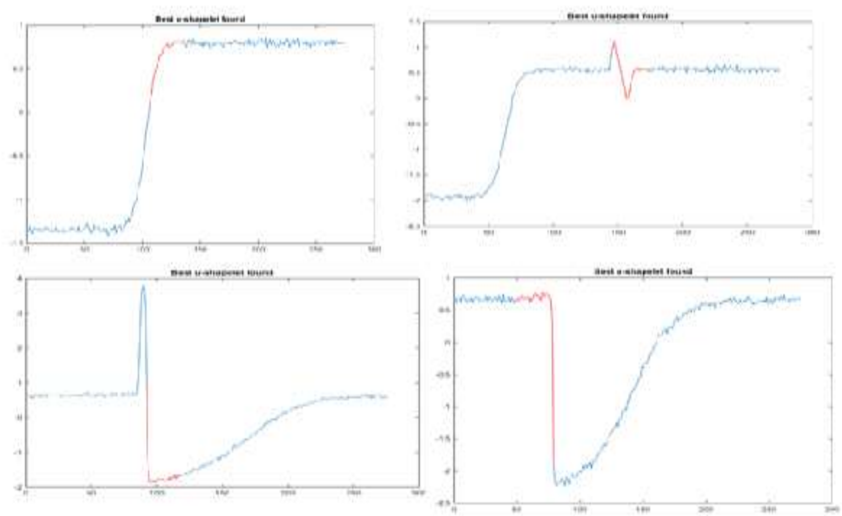

Figure 4: Best U-shapelets found in Trace Data Sample

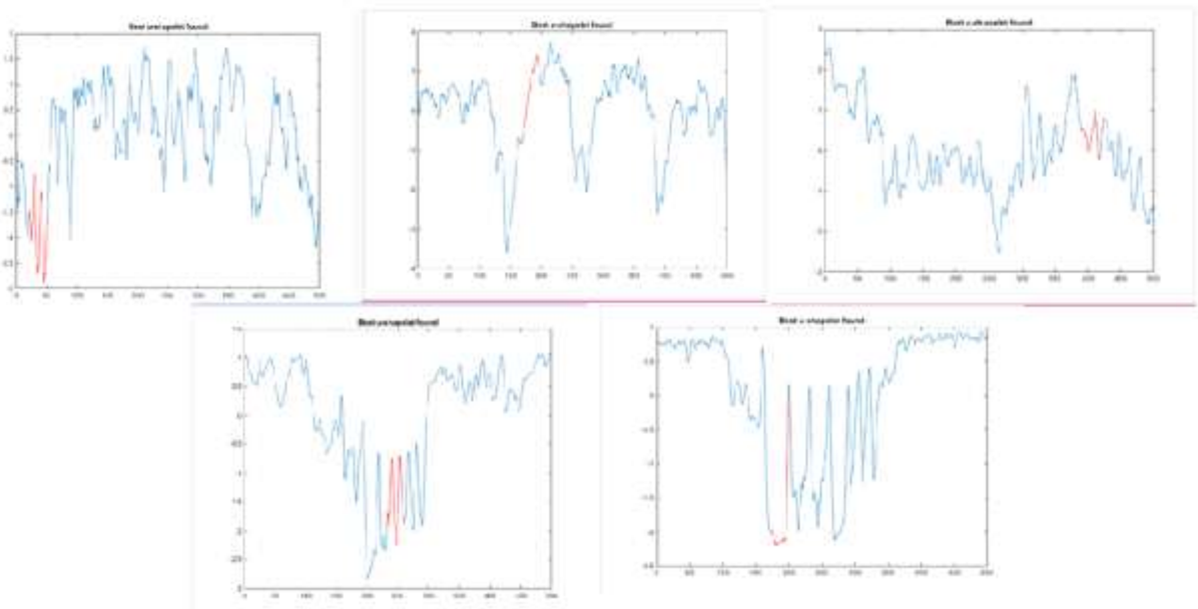

Figure5: Best U-Shapelets obtained for the Birds Data Sample

6. CONCLUSIONS

Clustering Time-Series is asignificantareainspired through numerous investigating demands comprising identical exploration in addition to challenges of generating approaches to distinguish dynamic changes in Time-Series. In this paper, a novel methodology is suggested to cluster the Time-Series data samples employing extracting u-shapelets from using the dynamic and efficient similarity measure. The dynamic time warping algorithm is employed as a similarity measure for extracting u-shapelets. This measure gave better accuracy and efficacy in forming the clusters when compared to the existing Euclidean distance measure. The DTW algorithm can be used for the data samples having different Time-Series lengths which are one of the advantages compared to ED measure. However, the computation time is more in the case of the DTW approach. The quality of the clustered data samples is measured using the Rand Index Approach and results showed that the DTW based TS clustering approach has a higher value compared to the ED measure.

REFERENCES

1. T.-Y. Chiu , T.-C. Hsu , C.-C. Yen , J.-S. Wang , Interpolation based consensus clustering for gene expression Time-Series, BMC Bio inform. 16 (117) (**2015**) 1–17 .
2. P. Cotofrei , K. Stoffel , Classification rules + time = temporal rules, Lecture Notes Comput. Sci. 23 (**2002**) 572–581

3. Ahmad Taufiq Hosni, Suharto Teriman, Nor Aizam Adnan, and Muhamad Asri Abdullah Kamar," Time Series Analysis of Land Cover Using Remote Sensing Technology", Journal of Computational and Theoretical Nanoscience ,17, 902–910 (**2020**).
4. B. DeVries, J. Verbesselt, L. Kooistra, M. Herold, Robust monitoring of small-scale forest disturbances in a tropical montane forest using Landsat Time-Series, Remote Sensing Environ. 161 (**2015**) 107–121.
5. G. Marti, S. Andler, F. Nielsen, P. Donnat, Clustering financial Time-Series: How long is enough? arXiv preprint arXiv:1603.04017 (**2016**) 1–7 .
6. T.W. Liao, Clustering of Time-Series data!' ªa survey, Pattern Recog. 38 (11) (**2005**) 1857–1874
7. S. Hirano and S. Tsumoto, "Cluster analysis of Time-series Medical data based on the Trajectory representation and multiscale comparison techniques," Proc. ICDM **2006**.
8. Robert H. Shumway, David S. Stoffer, "Time-Series Analysis and Its Applications", Springer Text in Statistics, **2011**.
9. Ye and Eamonn Keogh, "Time-Series Shapelets: A New Primitive for Data Mining," KDD'09, June 29–July 1, **2009**.
10. J. Lines, L. Davis, J. Hills, and A. Bagnall. A Shapeletstransforms for Time-Series classification. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **2012**.
11. A. Mueen, E. Keogh, and N. Young. Logical-Shapelets: an expressive primitive for Time-Series classification. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **2011**.
12. J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall. Classification of Time-Series by Shapelets transformation. Data Mining and Knowledge Discovery, **2013**.
13. J. Lines and A. Bagnall. Alternative quality measures for Time-Series Shapelets. In Intelligent Data Engineering and Automated Learning, volume 7435 of Lecture Notes in Computer Science, pages 475–483. **2012**.
14. T. Rakthanmanon and E. Keogh. Fast Shapelets: A scalable algorithm for discovering Time-Series Shapelets. Proceedings of the 13th SIAM International Conference on Data Mining, **2013**.
15. K.-W. Chang, B. Deka, W. mei W. Hwu, and D. Roth. Efficient pattern-based Time-Series classification on GPU. In M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. I. Webb, and X. Wu, editors, ICDM, pages 131–140. IEEE Computer Society, **2012**.
16. V. Niennattrakul; C.A. Ratanamahatana, "On Clustering Multimedia Time-Series Data Using K-Means and Dynamic Time Warping", International Conference on Multimedia and Ubiquitous Engineering, pp.733-738, April **2007**.
17. Shaozhi Wu, Yue Wu, Ying Wang, Yalan Ye, "An Algorithm for Time-Series Data Mining Based on Cluster", International Conference on Communications, Circuits, and Systems, IEEE, Vol. 3, pp. 2155-2158, June **2006**.

18. John Paparrizos, Luis Gravano, “k-Shape: Efficient and Accurate Clustering of Time-Series”, In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 1855-1870, **2015**.
19. S. Ongwattanakul; D. Srisai; “Contrast Enhanced Dynamic Time Warping Distance for Time-Series Shape Averaging Classification” ICIS 2009, November 24-26, **2009** Seoul, Korea.
20. X. Zhang, J. Liu, Y. Du, T. Lv, “A novel clustering method on Time-Series data”, Expert Systems with Applications, Science Direct, Vol. 38, No. 9, pp. 11891–11900, **2011**.
21. Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, Shaokai Wang, “Time-Series k -means: A new k -means type smooth subspace clustering for Time-Series data”, Information Sciences, Elsevier, Vol. 367, pp. 1-13, **2016**.
22. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. **2008**. Querying and Mining of Time-Series Data: Experimental Comparison of Representations and Distance Measures. In Proc of the 34th VLDB. 1542–1552.
23. J. Zakaria, A. Mueen, E. Keogh, Clustering Time-Series using Unsupervised-Shapelets, Proceedings of the IEEE International Conference on Data Mining, **2012**, pp. 785-794
24. Zakaria, J. (**2013**). Developing Efficient Algorithms for Data Mining Large Scale High Dimensional Data. *UC Riverside*. ProQuest ID: Zakaria_ucr_0032D_11593. Merritt ID: ark:/13030/m53218wb. Retrieved from https://escholarship.org/uc/item/660316zp.
25. J. R. Annam, S. S. Mittapalli and R. S. Bapi, "Time series Clustering and Analysis of ECG heart-beats using Dynamic Time Warping," *2011 Annual IEEE India Conference*, , pp. 1-3, Hyderabad, **2011.**
26. G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” Communication in ACM (**1975**) vol. 19, pp. 613–620.
27. supporting page: http://sites.google.com/site/ushapelet/
28. Carmela Iorio, Gianluca Frasso, Antonio D Ambrosio, Roberta Siciliano, "Parsimonious Time-Series clustering using P-splines", Expert systems with Applications, Elsevier, Vol. 52, pp. 26–38, **2016**.