An Alternate Measure for Comparing Time Series Subsequence Clusters

Ricardo Mardales **mardales@ engr.uconn.edu**

Dina Goldin **dqg@engr.uconn.edu**

BECAT/CSE Technical Report
University of Connecticut

1. Introduction

A recent paper [KLT03] claims that clustering of time series subsequences is a *meaningless* process. The basis for this claim lies on empirical observations that the result of clustering these subsequences is seemingly independent of the input. That is, when measuring the distance between the *cluster sets* for same time series, obtained by using different random seeds in the clustering algorithm, it is on the average no smaller than for two sets of clusters from different time series.

The problem with STS clustering is not that it is meaningless, but that *cluster set distance* is an inappropriate distance metric. The purpose of this study is to present and describe an alternate measure between clusters; that compares distances between *cluster shapes* rather than between cluster sets. Cluster shapes are composed of pair-wise distances between the cluster centers. With this new measure, we find that the input and output of subsequence clustering are no longer independent. In fact, our experiments show that given a cluster set, we can use its shape to correctly identify the series that produced it.

We begin by describing the process of Subsequence Time Series (STS) clustering. We then define the notion of *clustering meaningfulness* and review the work of [KLT03] which used the distance between cluster centers of various sets and failed to show its meaningfulness. Next, we present *cluster shapes*, pair-wise distances between the cluster centers in a set, as an alternate distance measure. *Cluster constellations* are the result of averaging together many cluster shapes for the same series. Then we describe our tests confirming that cluster constellations serve as good "fingerprint" for time series sequences, allowing us to reliably match cluster sets back to original sequences.

2. The Subsequence Clustering Process

As in [KLT03], we begin with a Time Series of length *m*. With a window size *w*, we slice the original time series into *m-w+1* subsequences, each of length *w* and we put these into a subsequence matrix M[m-w+1][w]. We then *normalize* each subsequence using the concept of *normal sequenc*e from [GK95]. Normalization is performed by subtracting the subsequence average from each coefficient and dividing it by the standard deviation; as a result each subsequence has an average of 0 and a standard deviation of 1.

Then, we cluster the normalized subsequences from M using the *k-means clustering algorithm* introduced in [Mac67], as follows:

1. We randomly choose *k* subsequences to be used as the first cluster centers. Their subsequence indexes serve as *seeds* into the algorithm, and can take values from 1 to (m-w+1).
2. For each subsequence in M, we calculate its *Euclidean Distance* to the *k* cluster centers and we assign it a value in the range [1, *k*] corresponding to the closest cluster center.
3. After each subsequence has been assigned to a cluster, we recalculate the cluster centers, taking the average of all subsequences members of that cluster.
4. We reassign all subsequences to the new cluster center with the minimum distance.
5. We repeat steps 2-4 until the cluster centers remain unchanged. A matrix C[k][w], which contains the cluster centers, is returned as the result.

We will refer to this process as a *clustering run*. Given a data sequence *S*, the number of clusters *k*, the window length *w*, and *k* initial seeds, a clustering run returns a unique set of cluster centers for *S*.

3. The Definition of Meaningful Clustering

In this section, we examine the concept of *clustering meaningfulness*. This concept is based on the observations from [KLT03] that "useful algorithms produce output that depends on the input" and that STS clustering "is meaningless if the output is independent of the input":

> **Definition 1 (clustering meaningfulness).** *A clustering algorithm is* meaningful *if its output is dependent on the input.*

More specifically, we define a *clustering meaningfulness test* as follows:

> **Definition 2 (clustering meaningfulness test).** *Given 10 time series from diverse domains, and ten sets of cluster centers, can one map the cluster centers back to the original time series?*

[KLT03] failed to show that clustering is meaningful. This work used *cluster set distance*, which is the sum of minimum Euclidean distances between cluster center sets resulting from different clustering runs.

Figure 1 illustrates the cluster set distance between two sets of 3 clusters X and Y; the cluster centers for X and Y are A,B,C and D,E,F, respectively. The cluster set distance between X and Y is the sum of the minimum distances AD + BD + CD; the cluster distance between Y and X is BD + BF + CE. Note that this distance measure is not commutative

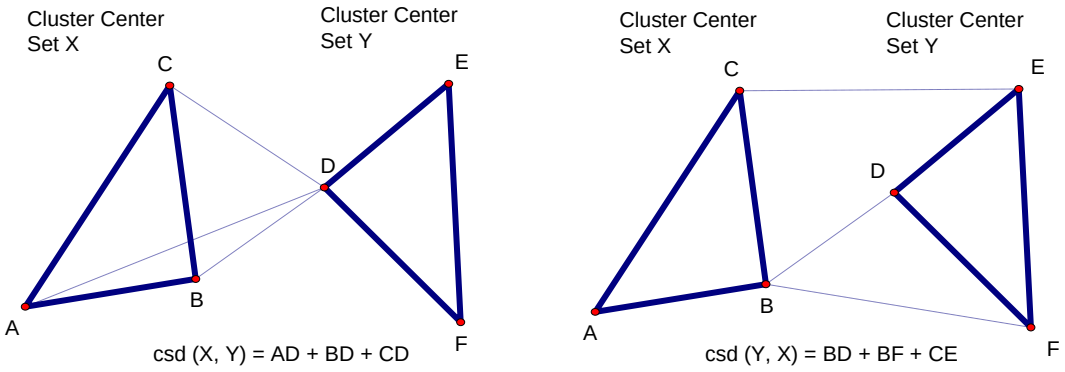

Figure 1: cluster set distance

Using cluster distances, [KLT03] measures "cluster meaningfulness" as follows:

1. First, they find the average cluster set distance (*csd*) for different clustering runs on the same series *S1*. For example, if there are three cluster sets (X1, X2, X3) then the distances are {*csd(X1, X2), csd(X1, X3), csd(X2, X3)*}.
2. Second, they calculate the same average cluster set distance, but this time for clustering runs on different series (*S1* and *S2*). For example, if there are three cluster sets (X1, X2, X3) from S1 and three cluster sets (Y1, Y2, Y3) from S2, then the distances are {*csd(X1,Y1), csd(X1,Y2), csd(X1,Y3)…*}.
3. Last, the ratio between the averages from step one and two is computed, as shown in (1). This *cluster meaningfulness ratio* serves as a measure of the meaningfulness of the STS clustering process.

$$\text{Cluster Meaningfulness (S1,S2)} = \frac{\text{average cluster set distance (S1,S1)}}{\text{average cluster set distance (S1,S2)}} \qquad (1)$$

Note that clustering meaningfulness compares two sequences at a time; if clustering is meaningful (Definition 1), then the cluster meaningfulness ratio should be in theory a value close to zero. This claim is based on the assumption that the cluster distance between cluster sets of the same series should be near zero, because that the sets would be similar to each other. Although [KLT03] claims that the "exact definition of clustering meaningfulness is not important", they do rely on their following assumption:

> **CSD assumption**: STS clustering is "meaningful" only if the cluster set distance for clustering runs over the same sequence is small.

As [KLT03] shows empirically, the cluster meaningfulness ratio is far from zero and closer to one in some cases. Given the empirical fact that there is not a considerable similarity between cluster centers regardless of the series used to produce them, [KLT03] goes on to conclude that subsequence series clustering does not produce meaningful results and that it fails the clustering meaningfulness test of Definition 2. In particular, they claim that "no one could tell the difference" if the cluster centers are the product of one series or the other.

We have implemented the subsequence clustering process and confirmed the results observed in [KLT03] using both Matlab and Java. We agree when measuring the cluster distance between the two sets of clusters for same time series, obtained by using different random seeds in the clustering algorithm, it is no smaller than for two sets of clusters from different time series as defined in the clustering meaningfulness ratio. In addition, we have confirmed that the weighted mean by cluster membership of the cluster center is a line or very close to one and the plot of these centers resemble sine waves.

However, we disagree with the conclusion that subsequence clustering is meaningless. Instead, we claim that it is the *cluster set distance* (csd) measure which does not produce meaningful results, and we conclude that *the CSD assumption is false*. We offer a more appropriate distance measure in the next section.

4. Cluster Shape Distance

In this section, we describe a new finding that led us to discover a different comparison approach that is not meaningless. In fact, with this new approach, given *N* series from different sources and *N* sets of cluster centers, these sets can be mapped back to their original series with high accuracy.

After clustering a set of time-series subsequences using the algorithm in Section 2, we obtain *k* cluster centers of length *w*. Then, we calculate the Euclidean distances for all pairs of cluster centers. For example, when k=3 we obtain 3 distances *D12, D13, D23*, where *Dij* represents the Euclidean distance between cluster centers *i* and *j*.

w	seeds (k=3)	*D12*	*D13*	*D23*	Sum	Average
8	[226;549;82]	2.7771	4.4644	2.4942	9.7357	3.245233
8	[902;7;171]	4.4855	2.7416	2.9164	10.1435	3.381167
8	[525;751;820]	4.4928	2.5958	2.8388	9.9274	3.309133
16	[226;549;82]	5.1477	6.5741	3.1317	14.8535	4.951167
16	[902;7;171]	6.6168	3.6607	4.8998	15.1773	5.0591
16	[525;751;820]	6.5801	3.2478	5.0325	14.8604	4.953467
32	[226;549;82]	8.3883	9.2677	3.7964	21.4524	7.1508
32	[902;7;171]	6.9156	9.4574	5.9889	22.3619	7.453967
32	[525;751;820]	9.2988	4.9502	7.4518	21.7008	7.2336

Table 1.

Table 1 illustrates the results for 9 different clustering runs over the ocean series from [KF02]. It shows the distances between cluster centers, as well as their sum and average. These results are the product of clustering with k=3 and w={8,16,32}; three different seeds are used for each combination of k and w.

It is easy to see that in Table 1, for any given combination of k and w, the sum and average are very close regardless of the initial choice of cluster centers. Furthermore, if we sort the set of distances *Dij*, the resulting lists of numbers {δ1, δ2, δ3} also similar for different initial seeds, as shown in Table 2.

W	seeds (k=3)	δ1	δ2	δ3	Sum	Average
8	[226;549;82]	2.4942	2.7771	4.4644	9.7357	3.245233
8	[902;7;171]	2.7416	2.9164	4.4855	10.1435	3.381167
8	[525;751;820]	2.5958	2.8388	4.4928	9.9274	3.309133
16	[226;549;82]	3.1317	5.1477	6.5741	14.8535	4.951167
16	[902;7;171]	3.6607	4.8998	6.6168	15.1773	5.0591
16	[525;751;820]	3.2478	5.0325	6.5801	14.8604	4.953467
32	[226;549;82]	3.7964	8.3883	9.2677	21.4524	7.1508
32	[902;7;171]	5.9889	6.9156	9.4574	22.3619	7.453967
32	[525;751;820]	4.9502	7.4518	9.2988	21.7008	7.2336

Table 2.

We will refer to this list of numbers as the shape of the cluster set:

Definition 3 (Cluster Shape). Given a set *S* of *k* clusters with centers {*C1, … Ck*}, the *shape of S* is the sorted sequence of *k*(k-1) / 2* numbers, representing the pair-wise Euclidean distances between the cluster centers {*C1, … Ck*}. It is denoted by {δ1, δ2, δ3...}.

As illustrated in Table 2, for a given series with any given combination of *k* and *w*, these shapes for various clustering runs tend to be similar to each other. The same effect occurred when we tried different series, or varied *k* and *w*.

This characteristic leads us to believe in the existence of some structure that determines the relative positions of the cluster centers in space, and that undergoes translations or rotations when different cluster seeds are chosen. As a result, while the absolute positions of cluster centers change and appear "meaningless", the structure itself does not change. We call this structure *cluster constellation,* as it describes the relative positions of cluster centers when viewed as points in multidimensional space.

Cluster constellations are simply the result of averaging together many cluster shapes for the same series:

Definition 4 (Cluster Constellation): A c*luster constellation* is the average of cluster shapes resulting from multiple clustering runs over the same series with the same *w* and *k*. Cluster constellations are denoted by {Δ1, Δ2, Δ3…}.

Figure 2 shows a cluster constellation (Δ1, Δ2, Δ3) for *k*=3, as the average of similar cluster shapes A, B, and C, created by three clustering runs. Let {a1,a2,a3} bethe *shape* of A; a1 is the shortest length between any two cluster centers in A, a2 is the second shortest, and a3 is the longest. Similarly, we compute {b1,b2,b3} and {c1,c2,c3}, the shapes of B and C. Then, Δ1 = avg(a1,b1,c1); Δ2 = avg(a2,b2,c2); Δ3 = avg(a3,b3,c3)).

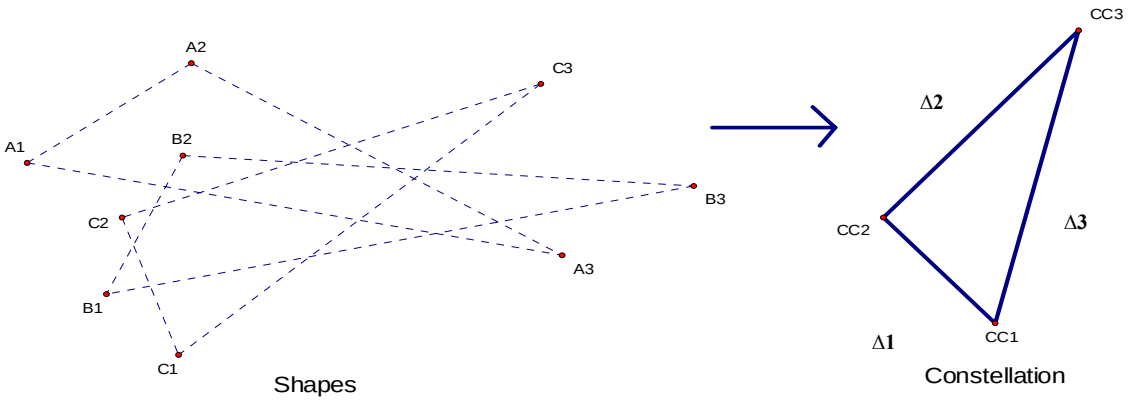

Figure 2: cluster constellation

5. How meaningful are cluster constellations?

When we compare cluster shapes to constellations, we see that the distance between them is near zero for the same series. Using the distances between cluster shapes (and constellations), as opposed to the distance between cluster sets (*cluster set distance* as described in part 3), we show that meaningful results can be obtained. In particular, we prove that STS clustering passes the *clustering meaningfulness test* of Definition 2.

Algorithm 1: Testing the Meaningfulness of Cluster Constellations.

1. Set a fixed *k* and *w*, which is applied throughout the process.
2. Calculate the cluster constellation (find the average of *Q* several shapes) for *N* different data series. Create a *master table* of these constellations.
3. For each of the *N* series from step 2, perform *R* clustering runs with various seeds, and create a *sample table* containing the resulting cluster shapes.
4. For each of the shapes in the sample table of each series, compute the Euclidean distance to each of the constellation fingerprints in the master table. Assign each shape to the series whose constellation in the master table is the closest. If the closest constellation is the one for the same series, then the assignment is *correct*.
5. For each series, compute the percentage of shapes in its sample table that were assigned correctly. Ideally, the result should be close to 100%; we will refer to it as the *meaningfulness score*.

To test that cluster constellations are not meaningless, we performed an experiment according to Algorithm 1. In our experiment, N=10, Q=100, and R=100. That is, we used 10 different data series from [KF02], averaging 100 cluster shapes to compute each constellation in the master table, and storing 100 shapes in each sample table.

An example master table with cluster constellation for five sequences from [KF02] is shown in Table 3, with *k=3* and *w=8*. Table 4 shows another master table for the same sequences, with *k=3* and *w=16*. This illustrates the relation between constellations and dimensionality; both increase as the value of *w* increases.

data	Δ1	Δ2	Δ3
ocean	2.3598	3.0464	4.4583
packet	2.1712	2.2315	2.3619
soil	1.9434	2.0073	2.0582
sp	2.5302	2.9574	3.774
tide	2.7175	3.3878	3.705

Table 3: *k=3*, *w=8*

data	Δ1	Δ2	Δ3
ocean	3.3018	5.1779	6.5902
packet	2.3881	2.4878	2.6392
soil	2.0046	2.3572	2.5495
sp	3.624	4.121	5.5512
tide	3.7356	4.2792	4.6382

Table 4: *k=3*, *w=16*

Table 5 is another master table, with *k=4* and *w=8*.

data	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6
ocean	2.4787	2.5844	2.9337	2.9778	3.2012	4.7337
packet	2.2235	2.2518	2.3438	2.5314	2.5647	2.6129
soil	2.0568	2.0874	2.1464	2.1803	2.2105	2.2533
sp	2.1239	2.465	2.9448	3.17	3.4565	4.1346
tide	2.4059	2.4585	3.3746	3.4247	4.0473	4.1776

Table 5: *k=4*, *w=8*

Then, we computed cluster shapes [δ1, δ2, δ3] for various series, and hid the information about the sequence that produced it. This is illustrated in Table 6. Each of the shapes that appears in Table 6 came from a different clustering run, and got assigned to some constellation from the master table on Table 3. We used the minimum Euclidean distance between shapes and constellations to assign a series. The assignments appear in the right-most column and all are correct.

δ1	δ2	δ3	Assignment
2.6517	2.9498	3.7824	sp
2.5873	3.5066	3.6869	tide
2.5958	2.8388	4.4928	ocean
2.1594	2.246	2.3478	packet
1.9323	2.0474	2.0711	soil
2.196	2.264	2.3352	packet
2.4942	2.7771	4.4644	ocean
2.5529	2.8036	3.7939	sp
1.9481	2.0417	2.0672	soil
2.8821	3.1982	3.7473	tide

Table 6: sample shapes and their assignment

We used 10 sequences from [KF02], and we repeated Algorithm 1 with different combinations of *w* and *k*. The results appear on Table 7. Overall there is an accuracy average of over 90%. This means that, given a set of cluster centers for any series, its shape will be correctly assigned to the original constellation and therefore we will correctly identify which series produced it.

k	**3**				**4**				**5**			
w	**8**	**16**	**32**	**64**	**8**	**16**	**32**	**64**	**8**	**16**	**32**	**64**
leleccum	100	100	100	83	97	92	91	99	98	93	100	83
robot	100	100	100	100	99	100	100	100	96	100	100	100
sensorA	100	100	100	100	100	100	100	100	98	100	100	96
burstin	89	100	100	100	89	100	100	100	97	100	100	100
infra	100	100	100	100	85	100	100	100	100	100	100	100
ocean	100	100	93	100	95	100	100	100	93	100	85	95
packet	100	100	87	100	100	100	100	100	100	100	100	100
soil	80	100	94	100	85	100	99	95	95	100	100	97
sp	90	100	100	100	75	100	100	100	98	90	100	100
tide	98	100	100	100	99	100	98	100	100	100	97	100

Table 7: meaningfulness scores

Our results in Table 7 prove that it is possible to tell which output is the result of its input, using the cluster shape distance measure. If we are able to identify the original series that produced the cluster centers, this proves that STS clustering is meaningful.

6. Discussion

The problem with STS clustering is not that it is meaningless, but that *cluster set distance* is an inappropriate distance metric. Instead we propose using *cluster shape distance*. While [KLT03] showed that clustering subsequence series is meaningless when cluster set distances are used, and the output is not related to the input, we have shown that the contrary is true when cluster shape distance is used.

Consider cluster meaningfulness ratio (Section 3); it was expected to be close to 0, but was found to be close to 1. We can conclude that if we use cluster shape distance instead, the cluster meaningfulness ratio will be close to zero:

$$\text{Cluster Meaningfulness (X,Y)} = \frac{\text{Cluster shape distance (X,X)}}{\text{Cluster shape distance (X,Y)}} \qquad (2)$$

Rather than looking at *absolute* positions of the centers, as measured by cluster set distance, we are only concerned with their *relative* positions, as measured by cluster shape distance. If we measure the distance between cluster sets, we create a dependency on the location of those sets. On the other hand, if we use shape distances, we obtain a measure that is independent of transformations, such as translations and rotations. We believe that this is the key difference between the two approaches.

The success of the approach based on relative, rather than absolute, positions, leads us to believe in the existence of some structure that determines the relative positions of the cluster centers in space, and that undergoes translations or rotations when different cluster seeds are chosen. As a result, while the absolute positions of cluster centers change and appear "meaningless", the structure itself does not change. We called this structure *cluster constellation,* as it describes the relative positions of cluster centers when viewed as points in multidimensional space.

Our work, just as [KLT03], is completely empirical. It is easily reproducible, and we look forward to having it confirmed by others. The challenge remains to explain these observations analytically. It is an open question why the locations of cluster centers change from one clustering run to another, while the cluster shape does not change.

Acknowledgment

We thank Eamonn Keogh for the correspondence regarding the notion of cluster meaningfulness.

Bibliography

[C84] Chatfield, C. The Analysis of Time Series: An Introduction, 3rd Edition. Chapman and Hall. New York (1984).

[D98] Das, G. et al. Rule Discovery from Time Series. In *Proceedings of the 4th Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 16-22. New York (1998).

[GK95] D.Q. Goldin, P.C. Kanellakis. On Similarity Queries for Time-Series Data: Constraint Specification and Implementation. 1st Int'l Conf. on the *Principles and Practice of Constraint Programming*, LNCS 976, pp. 137-153, Cassis France, Sep. 1995.

[HK01] Han, J., Kamber, M. Data Mining: Concepts and Techniques. Morgan Kauffman Publishers (2001).

[K76] Kendall, M. Time Series. Hafner Press (1976).

[KF02] Keogh, E. & Folias, T. The UCR Time Series Data Mining Archive. (2002) [http://www.cs.ucr.edu/~eamonn/TSDMA/index.html]. Riverside, CA. University of California-Computer Science and Engineering Department.

[KK02] Keogh, E. & Kasetty S. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. In *Proc. 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining,* pp.102-111. Edmonton, Canada (2002).

[KLT03] Keogh, E., Lin, J., Truppel, W. Clustering of Time Series is meaningless. *In Proc. IEEE Conf. on Data Mining, IEEE Computer Society* (2003) 115-122

[Mac67] MacQueen, J.: Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Math Statist*. Prob., 281-297, (1967).

[O99] Oates, T. Identifying Distinctive Subsequences in Multivariate Time Series by Clustering. In *Proc. of the 5th Int'l Conf. on Knowledge Discovery and Data Mining*. pp. 322-326, San Diego CA (1999).