Shape-Based Clustering for Time Series Data

Warissara Meesrikamolkul, Vit Niennattrakul,
and Chotirat Ann Ratanamahatana

Department of Computer Engineering, Chulalongkorn University
254 Phayathai Road, Pathumwan, Bangkok, Thailand 10330
{g53wms,g49vnn,ann}@cp.eng.chula.ac.th

Abstract. One of the most famous algorithms for time series data clustering is k-means clustering with Euclidean distance as a similarity measure. However, many recent works have shown that Dynamic Time Warping (DTW) distance measure is more suitable for most time series data mining tasks due to its much improved alignment based on shape. Unfortunately, k-means clustering with DTW distance is still not practical since the current averaging functions fail to preserve characteristics of time series data within the cluster. Recently, Shape-based Template Matching Framework (STMF) has been proposed to discover a cluster representative of time series data. However, STMF is very computationally expensive. In this paper, we propose a Shape-based Clustering for Time Series (SCTS) using a novel averaging method called Ranking Shape-based Template Matching Framework (RSTMF), which can average a group of time series effectively but take as much as 400 times less computational time than that of STMF. In addition, our method outperforms other well-known clustering techniques in terms of accuracy and criterion based on known ground truth.

Keywords: Time Series, Clustering, Shape-based Averaging.

1 Introduction

Time series data mining is increasingly an active research area since time series data are ubiquitous, appearing in various domains including medicine [15], geology [13], etc. One of its main mining tasks is clustering, which is a method to seperate unlabeled data into their natural groupings. In many applications related to time series data [14], k-means clustering [2] is generally used with the Euclidean distance function and amplitude averaging (arithmetic mean) as an averaging method.

Although the Euclidean distance is popular and simple, it is not suitable for time series data because its distance between two sequences is calculated in one-to-one manner. As a result, k-means with Euclidean distance does not cluster well because time shifting among data sequences in the same class usually occurs. In time series mining, especially in time series classification, Dynamic Time Warping (DTW) [1] distance has been proved to give more accurate results than Euclidean distance. Unfortunately, k-means clustering with the DTW

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 530–541, 2012.

distance still does not work practically [8][7] because current averaging function does not return a characteristic-preserving averaging result. Traditional k-means clustering fails to return a correct clustering result since this cluster centers do not reflect characteristics of the data, as shown in Fig. 1. In this work, we will demonstrate that our proposed method can resolve this problem.

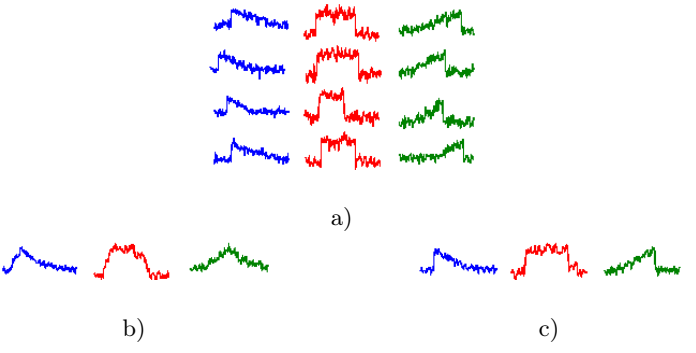


Fig. 1. a) Sample 3-class CBF data [3] and its cluster centers from b) traditional k-means clustering and from c) our proposed method

We propose a novel method called Shape-based Clustering for Time Series (SCTS) which incorporates k-means clustering and DTW distance measure, together with our new averaging method, called Ranking Shape-based Template Matching Framework (RSTMF) extended from Shape-based Template Matching Framework (STMF) [10] for classification. Unlike STMF, our RSTMF uses distances from clustering to approximate an order of sequences to be averaged, giving a few orders of magnitude speedup comparing to STMF. Our evaluation also shows that our proposed method outperforms other well-known clustering techniques in terms of accuracy and criterion based on known ground truth. In addition, the accuracy of our proposed method can future improve when a global constraint [11] is utilized in distance calculation and data averaging.

The rest of the paper is organized as follows. In section 2 and 3, we offer background knowledge and related works. In section 4, we explain our new framework for time series clustering, which is Shape-based Clustering for Time Series (SCTS). The experiments and results are shown in section 5. Finally, conclusions are provided in section 6.

2 Background

This section provides background knowledge on k-means clustering, Dynamic Time Warping (DTW) distance measure, and global constraint.

2.1 *K*-means Clustering

K-means clustering [2] is a well-known and very simple partitioning clustering algorithm. Its algorithm tries to group similar data into the same cluster by using an objective function that minimizes a sum of squared errors between a cluster center to its members. The algorithm is done as follows:

1. Initialize k cluster centers.
2. Measure the similarity between each data and all cluster centers and assign data into the most similar cluster.
3. Calculate a new cluster center of every cluster using an averaging function.
4. Repeat steps 2 and 3 until the cluster membership does not change.

K-means clustering consists of two major subroutines, which are a distance function to measure the similarity between data sequences and an averaging function to return a new cluster center. Generally, most time series clustering works use Euclidean distance and amplitude averaging method. However, both cluster centers and their cluster members are inaccurate. In this work, we resolve this problem by using the DTW distance measure with our newly proposed averaging method called RSTMF.

2.2 Dynamic Time Warping (DTW) Distance Measure

DTW distance [1] is an accurate similarity measurement which is generally used for time series data [9], especially in classification [6]. An optimal alignment and distance between two sequences $P = \langle p_1, \ldots, p_i, \ldots, p_n \rangle$ and $Q = \langle q_1, \ldots, q_j, \ldots, q_m \rangle$ can be determined as follows.

$$DTW(P,Q) = \sqrt{dist(p_n, q_m)} \tag{1}$$

$$dist(p_i, q_j) = (p_i - q_j)^2 + min \begin{cases} dist(p_{i-1}, q_j) \\ dist(p_i, q_{j-1}) \\ dist(p_{i-1}, q_{j-1}) \end{cases} \tag{2}$$

DTW distance is computed through dynamic programming to discover the minimum cumulative distance of each element in $n \times m$ matrix. In addition, the warping path between two sequences can be found by tracing back from the last cell.

In this work, DTW distance is used to measure the similarity between each time series data and cluster centers to give more accurate results.

2.3 Global Constraint

The global constraint is used when we need to limit the amount of warping in the DTW alignment. In some applications such as speech recognition [12], two data sequences are considered the same class when only small time shifting occurs; so,

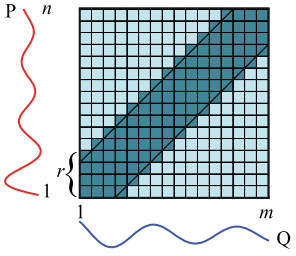


Fig. 2. The warping window of P and Q is limited by the global constraint of size r

the global constraint is used to align the sequences more precisely. The Sakoe-Chiba band [12], one of the most popular global constraints, has been originally proposed for speech community and also has been used in various tasks in time series mining [11]. The size of the warping window is defined by r (as shown in Fig. 2), the percentage of the time series' length, which is symmetric in both above and on the right of a diagonal. In this work, we will show in experiments that the global constraint plays an important role in improving the accuracy.

3 Related Work

In the past few decades, there are many clustering techniques proposed to cluster time series data [5], for example, agglomerative hierarchical clustering [13], which merges most similar objects until all objects are in the cluster. However, this technique is still inaccurate, especially when outliers are present.

Another popular clustering technique is partitional clustering, which tries to minimize an objective function. The well-known algorithms are k-medoids and k-means clustering, which are different in their approaches to find new cluster centers. For k-medoids clustering application [4], DTW distance is used as a similarity measure among data sequences, and a sequence with minimum sum of distance to the rest of the sequences in the cluster is selected as a new cluster center. However, medoid is not always a centroid of a cluster, so the sequences can be assigned to wrong clusters.

In contrast to k-medoids clustering, k-means clustering mostly uses Euclidean distance as a distance metric, and an arithmetic mean or amplitude averaging is simply used to find a new cluster center [14]. Although the DTW distance is more appropriate for time series data, there currently is no DTW averaging method that provides a satisfied averaging result.

According to this, many research works have tried to improve the quality of the averaging result. Shape-based Template Matching Framework (STMF) [10] was recently introduced to average time series sequences. Table 1 shows the algorithm of this framework; the most similar pair of sequences is averaged by Cubic-spline Dynamic Time Warping (CDTW) algorithm (in line 6).

Table 1. Shape-based Template Matching Framework algorithm [10]

Algorithm STMF(D)	
1.	D is the set of time series data to be averaged
2.	initialize weight $\omega = 1$ for every sequences in D
3.	while(size(D) > 1)
4.	$\{C_1, C_2\}$ = the most similar pair of sequences in D
5.	Z = CDTW(C_1, C_2, ω_{C_1}, ω_{C_2})
6.	$\omega_Z = \omega_{C_1} + \omega_{C_2}$
7.	add Z to D
8.	remove C_1, C_2 from D
9.	end while
10.	return Z

Given C_1 and C_2 as the most similar sequences, first, we find the warping path between these two sequences. The variables c_{1i} and c_{2j} are elements of C_1 and C_2, which are warped. The averaged sequence Z, which has coordinates z_{k_x} and z_{k_y} can be computed as follows.

$$z_{k_x} = \frac{\omega_{c_1} c_{1i} + \omega_{c_2} c_{2j}}{\omega_{c_1} + \omega_{c_2}} \tag{3}$$

$$z_{k_y} = \frac{\omega_{c_1} c_{1i_x} + \omega_{c_2} c_{2j_y}}{\omega_{c_1} + \omega_{c_2}} \tag{4}$$

In equations 3 and 4, ω_{c_1} and ω_{c_2} are the weight of the sequences C_1 and C_2, respectively. After we get the result, a number of points in the averaged sequence is re-sampled by using cubic-spline interpolation [10]. As shown in Fig. 3a), the averaging result from DTW averaging gives a sequence with 9 unequally spaced data points, whereas in Fig. 3b), the sequence is resampled with cubic spline interpolation to obtain a sequence of 7 equally spaced data points.

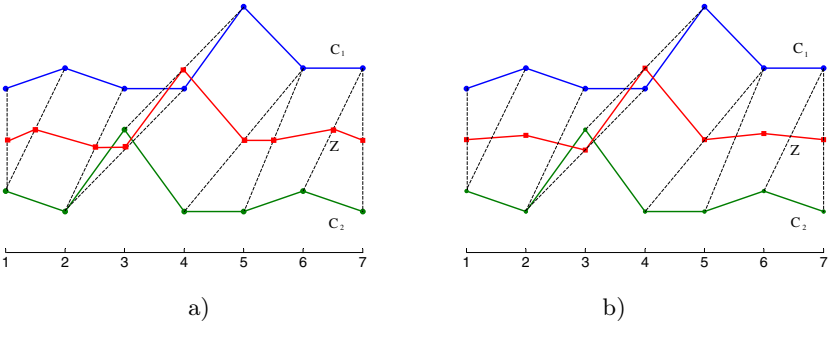


Fig. 3. The average sequences between C_1 and C_2 using DTW alignment a) before applying cubic spline interpolation and b) after applying cubic spline interpolation

However, according to this framework, finding the most similar pair for each time of averaging is enormously computationally expensive because the DTW distance of every pair of the sequences must be computed. Therefore, our RSTMF will mainly focus on improving its time complexity by estimating an order of sequences before averaging while maintaining the accuracy of the averaging results.

4 Shape-Based Clustering for Time Series (SCTS)

In this paper, we propose Shape-based Clustering for Time Series (SCTS) by incorporating k-means clustering and DTW distance, together with a novel averaging function, Ranking Shape-based Template Matching Framework (RSTMF). Although STMF can still be used to determine a cluster center, it is computationally expensive; therefore, computational time of k-means clustering significantly increase.

We provide an overview of the proposed clustering algorithm in Table 2; the DTW distance is used instead of the Euclidean distance in a membership assignment process. After we finished assigning each data sequence into the most similar cluster, RSTMF is utilized to average all of the sequences within each cluster until all cluster centers are updated. Unlike STMF, RSTMF approximates an order of averaged sequences by looking at the $Dist$ value, which is the DTW distance between data sequences in M and all cluster centers in C. Accordingly, RSTMF can provide the average sequence by using less computation time than that of STMF, which calculates the distance between every pair of data and the most similar pair of sequences is averaged, making it very computationally expensive.

Table 3 shows our RSTMF averaging algorithm, which determines a cluster center by using Cubic-spline Dynamic Time Warping (CDTW) [10] to average a pair of time series sequences. RSTMF utilizes $Dist$ to approximate a similarity distance between every sequence pair, defined by $dist_{approx}$. After that, CDTW is used to average a pair of sequences with the minimum $dist_{approx}$ value. Then, we update S and continue the averaging until only one sequence remains.

In RSTMF algorithm, the $dist_{approx}$ between each pair of the sequences can be computed by using the $Dist$ value. Suppose P and Q are data sequences in M, we have $Dist_{M_P,\ldots} = \langle Dist_{M_P,C_1}, \ldots, Dist_{M_P,C_k}, \ldots, Dist_{M_P,C_K} \rangle$ and $Dist_{M_Q,\ldots} = \langle Dist_{M_Q,C_1}, \ldots, Dist_{M_Q,C_k}, \ldots, Dist_{M_Q,C_K} \rangle$ where $Dist_{M_P,C_k}$ and $Dist_{M_Q,C_k}$ are the distance between P or Q and its k^{th} cluster center, and K is a number of cluster. By applying the triangular inequality theorem, p_k and q_k are assumed to be two sides of a triangle. Then, the $dist_{approx}$ of P and Q, which is another side of the triangle, can be approximated by equation 5 and collected into S.

$$dist_{approx}(Dist_{M_P,\ldots}, Dist_{M_Q,\ldots}) = \max_{1 \le k \le K} \left| Dist_{M_P,C_k} - Dist_{M_Q,C_k} \right| \quad (5)$$

After finishing an averaging of two sequences, we insert the resulting sequence into M and delete these two sequences. Then, we update S by using the algorithm in Table 4.

Table 2. Shape-based Clustering for Time Series (SCTS)

Algorithm SCTS(D, K)	
1.	D is the set of time series data
2.	C is the set of cluster centers
3.	K is the number of cluster in C
4.	M is the set of data in each cluster
5.	$Dist$ is the matrix of the distance between data sequences and all cluster centers
6.	initialize C as cluster centers of K clusters
7.	do
8.	for i = 1:size(D)
9.	for k = 1:K
10.	$Dist_{D_i,C_k}$ = DTW(D_i, C_k)
11.	end for
12.	if($Dist_{D_i,C_k}$ is minimal)
13.	assign D_i into M_k
14.	end if
15.	end for
16.	for k = 1:K
17.	C_k = RSTMF(M_k, $Dist$)
18.	end for
19.	while(the cluster membership changes)
20.	return the cluster members and the cluster centers

Table 3. The RSTMF algorithm

Algorithm RSTMF(M, $Dist$)	
1.	M is the set of data in each cluster
2.	$Dist$ is the matrix of the distance between data sequences and all cluster centers
3.	S is the matrix of the distance between data sequences in M
4.	initialize weight $\omega = 1$ for every sequences in M
5.	for i = 1:size(M)
6.	for j = i+1:size(M)
7.	$S_{M_i,C_j} = S_{M_j,C_i} = dist_{approx}(Dist_{M_i,\ldots}, Dist_{M_j,\ldots})$
8.	end for
9.	end for
10.	while(size(M) > 1)
11.	S_{M_i,C_j} = minimum value in S
12.	M_z = CDTW(M_i, M_j, ω_{M_i}, ω_{M_j})
13.	$\omega_{M_z} = \omega_{M_i} + \omega_{M_j}$
14.	add M_z to M
15.	UPDATE(S, i, j, z)
16.	remove M_i, M_j from M
17.	end while
18.	return M_z

Table 4. The UPDATE algorithm

Algorithm UPDATE(S, a, b, z)	
1.	S is the matrix of the distance between data sequences in M
2.	for i = 1:size(S)
3.	$S_{M_z,M_i} = S_{M_i,M_z} = \min(S_{M_a,M_i}, S_{M_b,M_i})$
4.	end for
5.	remove $S_{M_a,\ldots}$, $S_{\ldots,M_a}$, $S_{M_b,\ldots}$, $S_{\ldots,M_b}$ from S

By using the $dist_{approx}$ and the UPDATE method, our RSTMF can achieve large speedup because we can estimate an order of the sequences before averaging. In contrast, the original STMF needs to calculate the DTW distance to select the most similar pair of the sequences every time of averaging.

5 Experiments and Results

In this work, we evaluate our method by comparing it with other clustering techniques, which are typical k-means clustering with the Euclidean distance and amplitude averaging function, k-medoids clustering with the DTW distance [4], and k-hierarchical clustering [13] using both the Euclidean and the DTW distance. We compare our SCTS using RSTMF with that using the original STMF. Our experiments are evaluated on ten datasets from the UCR datasets classification/clustering archive [3] in diverse domains, as shown in Table 5.

Table 5. The details of datasets

Datasets	Number of classes	Length of data	Size of training set	Size of test set
Synthetic Control	6	60	300	300
Trace	4	275	100	100
Gunpoint	2	150	50	150
Lightning-2	2	637	60	61
Lightning-7	7	319	70	73
ECG	2	96	100	100
Olive Oil	4	570	30	30
Fish	7	463	175	175
CBF	3	128	30	900
Face Four	4	350	24	88

We execute each algorithm for 40 times with random initial cluster centers, and the k value is set to the a number of classes in each dataset. With the luxury of labeled datasets used in all experiments, an accuracy, which is the number of correctly assigned data sequences in all clusters, is used evaluation. Fig. 4 shows the accuracy of our proposed method, comparing other well-known clustering methods mentioned above. According to the results, our method outperforms others in almost all datasets.

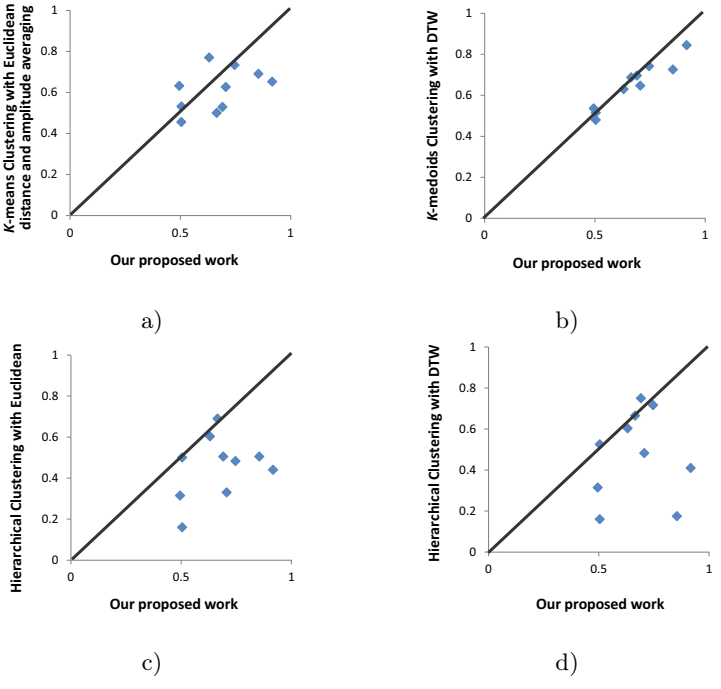


Fig. 4. The accuracy of our RSTMF method on 10 datasets, comparing with a) general k-means clustering, b) k-medoids clustering, and k-hierarchical clustering using c) the Euclidean distance and d) the DTW distance, respectively

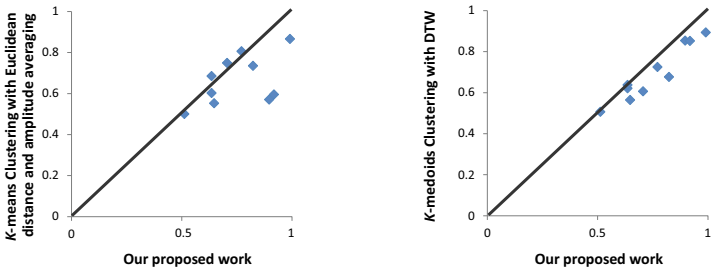
To re-emphasize our finding, we also use another criterion based on known ground truth [5] to measure a similarily between two sets of clusters, i.e., ground-truth clusters and results from clustering algorithms. Suppose G and C are sets of k ground truth clusters and the clusters from our clustering technique. The similarity between G and C is calculated by the following equations.

$$Sim(G, C) = \frac{1}{k} \sum_{i=1}^{k} \max_{1 \leq j \leq k} Sim(G_i, C_j) \tag{6}$$

$$Sim(G_i, C_j) = \frac{2\,|G_i \cap C_j|}{|G_i| + |C_j|} \tag{7}$$

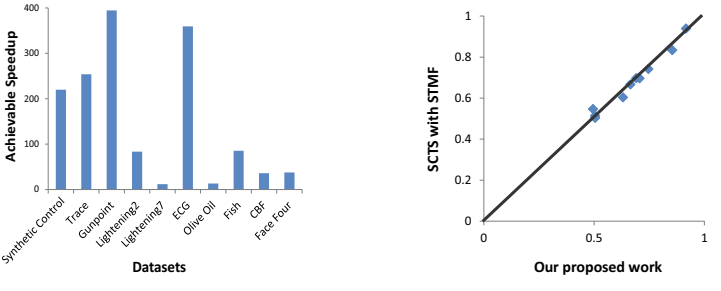
In Fig. 5, we compare our proposed work with the general k-means clustering and the k-medoids clustering using this criterion. The results show that the clusters obtained from our method are more similar to the ground-truth clusters because the RSTMF averaging method does give the new cluster centers that represent the overall charactheristic of the data within each cluster.

Furthermore, RSTMF can reduce the time complexity by a few orders of magnitude (as shown in Fig. 6a), while still providing comparable accuracy to STMF (as shown in Fig. 6b).



a) b)

Fig. 5. The criterion based on known ground ground truth, comparing our proposed method with a) general k-means clustering and b) k-medoids clustering



a) b)

Fig. 6. a) The speedup achieved by our proposed work. b) The accuracy of our proposed work comparing with that using STMF.

In some cases, it appears that SCTS with DTW distance achieves a lower accuracy than the general k-means clustering. In an attempt to alleviate this drawback, we experiment on the global constraint parameter of DTW, Sakoe-Chiba band. We can improve the clustering accuracy, comparing with the original k-means clustering (warping window size is 0%). Fig. 7 shows the accuracy of our proposed RSTMF and STMF, which are comparable, as warping window sizes vary. In almost datasets, the larger warping window size does not always provide the better accuracy; so, the appropriate warping window size is around 20%. However, in some dataset such as ECG, the wider warping window can lead to pathological warping and make the accuracy of clustering decreases.

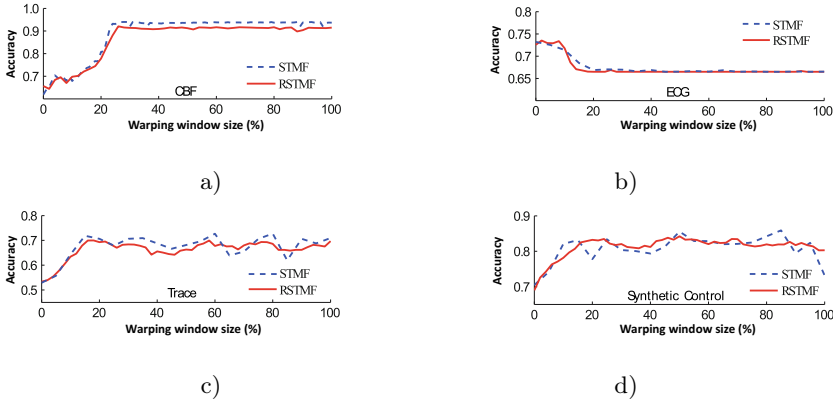


Fig. 7. The accuracy of Shape-based clustering using STMF and our proposed RSTMF of a) CBF, b) ECG, c) Trace, and d) Synthetic Control datasets

6 Conclusion

In this paper, we propose time series data clustering technique called Shape-based Clustering for Time Series (SCTS), which incorporates k-means clustering with a novel averaging method called Ranking Shape-based Template Matching Framework (RSTMF).

Comparing with the other well-known clustering algorithms, our SCTS yields better cluster results in terms of both accuracy and the criterion based on known ground truth because our RSTMF averaging function provides cluster centers that preserve characteristics of data sequences within the cluster (as shown in Fig. 8). Furthermore, RSTMF does gives a comparable sequence averaging result while consuming much less computational time than STMF in a few orders of magnitude; therefore, RSTMF is practically applied in clustering algorithm. We also used global constraint to increase an accuracy of our clusters. The results show that our SCTS can provide more accurate clustering when the width of warping window is about 20% of time series length.

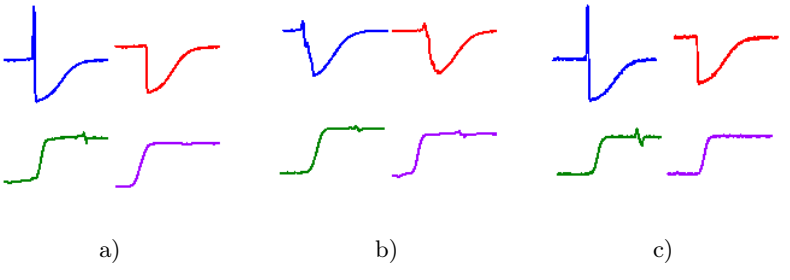


Fig. 8. The cluster centers obtained from a) our proposed method and b) the original k-means clustering of c) sample 4-class Trace data

Acknowledgements. This research is partially supported by the Thailand Research Fund (Grant No. MRG5380130), the Thailand Research Fund given through the Royal Golden Jubilee Ph.D. Program (PHD/0141/2549 to V. Niennattrakul and C.A. Ratanamahatana), and CU Graduate School Thesis Grant.

References

1. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Proceedings of AAAI Workshop on Knowledge Discovery in Databases, pp. 359–370 (1994)
2. Bradley, P.S., Fayyad, U.M.: Refining initial points for k-means clustering. In: Proceedings of the International Conference on Machine Learning (ICML 1998), pp. 91–99 (1998)
3. Keogh, E., Xi, X., Wei, L., Ratanamahatana, C.A. (2011), `http://www.cs.ucr.edu/~eamonn/time_series_data`
4. Liao, T.W., Bodt, B., Forester, J., Hansen, C., Heilman, E., Kaste, R.C., O'May, J.: Understanding and projecting battle states. In: Proceedings of 23rd Army Science Conference (2002)
5. Liao, T.W.: Clustering of time series data-a survey. Pattern Recognition, 1857–1874 (2005)
6. Meesrikamolkul, W., Niennattrakul, V., Ratanamahatana, C.A.: Multiple shape-based template matching for time series data. In: Proceedings of the 8th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2011), pp. 464–467 (2011)
7. Niennattrakul, V., Ratanamahatana, C.: On clustering multimedia time series data using k-means and dynamic time warping. In: Proceedings of the International Conference on Multimedia and Ubiquitous Engineering, pp. 733–738 (2007)
8. Niennattrakul, V., Ratanamahatana, C.A.: Inaccuracies of Shape Averaging Method Using Dynamic Time Warping for Time Series Data. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007. LNCS, vol. 4487, pp. 513–520. Springer, Heidelberg (2007)
9. Niennattrakul, V., Ruengronghirunya, P., Ratanamahatana, C.: Exact indexing for massive time series databases under time warping distance. Data Mining and Knowledge Discovery 21, 509–541 (2010)
10. Niennattrakul, V., Srisai, D., Ratanamahatana, C.A.: Shape-based template matching for time series data. Knowledge-Based Systems 26, 1–8 (2011)
11. Ratanamahatana, C.A., Keogh, E.: Making time-series classification more accurate using learned constraints. In: Proceedings of SIAM International Conference on Data Mining (SDM 2004), pp. 11–22 (2004)
12. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 43–49 (1978)
13. Shumway, R.H.: Time-frequency clustering and discriminant analysis. Statistics and Probability Letters, 307–314 (2003)
14. Vlachos, M., Lin, J., Keogh, E., Gunopulos, D.: A wavelet-based anytime algorithm for k-means clustering of time series. In: Proceedings of Workshop on Clustering High Dimensionality Data and Its Applications, pp. 23–30 (2003)
15. Wismuller, A., Lange, O., Dersch, D.R., Leinsinger, G.L., Hahn, K., Pütz, B., Auer, D.: Cluster analysis of biomedical image time-series. International Journal of Computer Vision, 103–128 (2002)