COMPARING EFFICIENCY BETWEEN TWO MEASURES OF EUCLID AND DTW USED IN DISCOVERY MOTIF IN TIME SERIES

NGUYEN TAI DU[1], PHAM VAN CHUNG[2]

[1]*Industry University of HCM city*
[2] *Faculty of Information Technology, Industry University of HCM city*
taidunguyen@gmail.com - *pchung@iuh.edu.vn*

Abstract. The study on time series databases which is based on efficient retrieval of unknown patterns and frequently encountered in time series, called *motif*, has attracted much attention from many researchers recently. These *motifs* are very useful for the exploration of data and provide solutions to many problems in various fields of applications. In this paper, we try to study and evaluate the efficiency of the use of both Euclidean and Dynamic Time Warping (DTW) distance meassures, utilizing Brute-force and Mueen – Keogh algorithms (MK), of which MK algorithm has performed efficiently in terms of CPU time and the accuracy of the problem of discovery the *motif* patterns. The efficiency of this method has been proven through experiment on real databases.

Keywords. time series, motif discovery, DTW measure, Euclidean measure, Sakoe Chiba limit, LB_Keogh limit.

1 INTRODUCTION

Currently, technology is constantly developing, the volume of data information is increasing rapidly in many areas such as science, technology, health, finance, economy, education, space. bioinformatics, robots ... Time series is a tuple of *m* real numbers measured at equal time intervals. They arise in many fields: internet, books, television, environment, stock, hydrometeorology, high tide ... This is a very useful resource to find useful information. Many researchers have used many methods to data mining on the time series for many years.

Discovery motifs in time series data has been used to solve problems in a variety of application areas since 2002, such as using motifs for signature verification [1], to detect for duplicate images in the shape database, to forecast stock prices [2], to classify time series data [3] and also be used as pre-processing in more advanced data mining operations.

The algorithm for identifying the exactly motifs (Brute-force) is quadratic in *n*, the number of individual time series (or the length of the single time series from which subsequences are extracted) [4]. To increase the time efficiency in identify motif. Some approximation algorithms have been proposed [5,6,7,8,9],. These algorithms have the cost of $O(n)$ or $O(n\log n)$, however, they require some predefined parameters.

Most algorithms of data mining in the time series need to compare time series by measuring the distance between them. Usually the Euclidean distance or DTW distance is used However, the Euclidean distance has been shown execution time much faster than DTW measurement but it is easy to break [10] [7]. DTW measurements have been used as a technique to allow for more accurate calculation of distances in case the time series has the same shape, but the number of points on them varies. In 2009 a new method introduced for data mining on time series and sequential data reduced the execution time when using DTW measurement [11]. The choice of using the measure affects the execution time and the accuracy of the results. In this paper, we use the discovery motif problem to compare and evaluate the effectiveness and execution time of two measures of Euclid and DTW.

In this work, we experimented by implementing two Brute-force algorithms and MK algorithms and using both measures of Euclid and DTW. In addition, we rely on the two ideas of J. Lin and Keogh, E., Pazzani [10] to introduce the extension of two Euclid and DTW measures combining the Piecewise Aggregate Approximation (PAA) number reduction technique in discovery motif problem on time series.

The rest of the paper is organized as follow. In section 2, we present some background knowledge about discovery motifs and distance mesurements, some methods of reducing the number of dimensions, discrete data. Section 3, compares DTW and Euclidean measurements on Brute force and MK algorithms. Section 4, Experimenting to evaluate the results of two MK motif mining algorithms and Brute-force algorithm on two distances measuring DTW and Euclidean. The rest of the paper gives some conclusions and future work.

2 BACKGROUND

In this section, we provide some background knowledge in the discovery motif based on calculating the distance measurement on subsequences in the time series

2.1 Definitions [4,12,]

Definition 1

Time Series: A time series $T = t_1, \dots , t_m$ is an ordered set of n real-valued variables (elements in the set can be repeated).

Definition 2

Subsequence: Given a time series T of length m, a subsequence C of T is a sampling of length $n < m$ of contiguous position from T, that is, $C = t_p, \dots ,t_{p+n-1}$ for $1 \leq p \leq m - n + 1$.

Definition 3

A *Time Series Database* (D) is an unordered set of m time series possibly of different lengths

Definition 4

The *Time Series Motif* of a time series database D is the unordered pair of time series $\{T_i, T_j\}$ in D which is the most similar among all possible pairs. More formally, $\forall a,b,i,j$ the pair $\{ T_i, T_j\}$ is the motif iff $\text{dist}(T_i, T_j) \leq \text{dist}(T_a, T_b)$, $i \neq j$ and $a \neq b$.

2.2 Motifs discovery

There are two main approaches in the discover motif:

The exact motif: the discover motif on the original data, based on the brute-force algorithm as a basis and thereby can improve the algorithm by applying a number of heuristic to accelerate and reduce the complexity of the algorithm.

However, these approach-based algorithms have high accuracy and completeness while runtime is not high. It is only suitable for small size data.

Approximate motif: time series data will be pre-processed before making mining such as reducing the number of dimensions, discrete data. During the mining process, some properties based on probability and randomness can be applied. This approach increases the effectiveness of algorithms while being correct and acceptable. It is suitable for large size data.

2.3 Similar distance measurement

For checking the two subsequences that they are a different or not, must be used a distance function. If the value of the distance function is zero, the two subsequences are the same. If the value of the distance function is greater, they are the more different. Two commonly used distance measurements are Euclidean and DTW

Euclidean distance

Euclidean distance is calculated by the following function with p = 1

$$\boldsymbol{D(Q,C) = \sqrt[p]{\sum_{i=1}^{n}(|Q_i - C_i|)^p}}$$

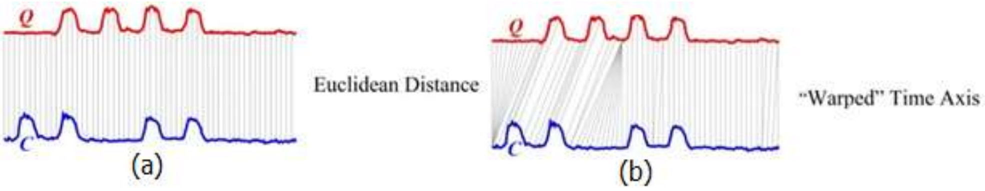


Figure 1: (a) The Euclidean distance of *Q* and *C*, (b) The Dynamic time warping distance of *Q* and *C* [11]

Dynamic Time Warping Distance (DTW)

In this case, a point from the *Q* can be mapped to multiple points in the *C* and these maps are not aligned. Then use DTW, Figure 1b illustrates this.

DTW gives more accurate results, but runtime is much higher than Euclidean.

2.4 Dimensional reduction and discrete on the original time series

The size of the time series data is often very large. Therefore, it needs to be transformed into shorter and simpler data by reducing the number of dimensions or reducing the size and discrete data into bits or characters to improve retrieval and computation efficiency.

Dimensional reduction

The dimensional reduction is the representation of the *n*-dimensional time series data $X = (x_1, \dots, x_n)$ into the *k*-dimensional lines $Y = (y_1, \dots, y_k)$, *Y* is called the baseline and *k* is the coefficient of the baseline. From the basic *Y*, the data can completely restore the initial *X* data.

Piecewise Aggregate Approximation method (PAA)

The Piecewise Aggregate Approximation method (PAA) proposed by E. Keogh et al. 2001 [13] as shown in Figure 2. This method approximates *k* points of contiguous values into the same mean value of *k* points. The process is done from left to right and the end result is a ladder line. Calculated time is very fast, supports queries with different lengths. However, rebuilding the initial sequence is very difficult, often producing errors and ignoring the extreme points in each approximation segment because of the mean value.

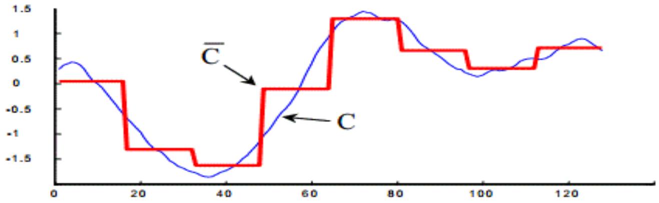


Figure 2: PAA method

Discrete data

The most commonly used discretization method is Symbolic Aggrigate Appriximation (SAX) that converts time series data into strings of characters. This method was proposed by J. Lin [14]. The original data was discretized by the PAA method, with each fragment in the PAA subsequence mapped to a corresponding letter based on the Gauss standard distribution as shown in Figure 3.

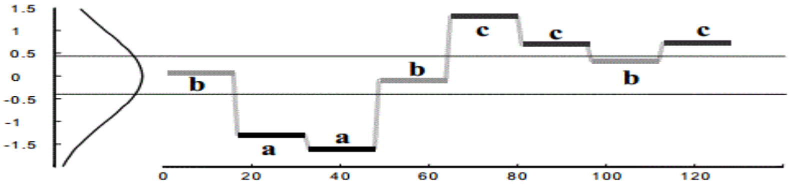


Figure 3: Symbolic Aggrigate Approximation method

SAX is suitable for characterizing data, it can interact with large data (Terabyte rows), suitable for data processing on the string, suitable for motif identification problems. However, breakpoints are defined based on the standard distribution (Gauss), which cannot be appropriate for all types of data.

3 COMPARING EFFECTIVENESS OF DTW AND EUCLIDEAN MEASUREMENTS IN BRUTE-FORCE AND MK ALGORITHMS

When studying algorithms, it is important to consider the runtime and algorithm efficiency (high accuracy results). A discovery motif algorithm is only considered to be optimal if it meets the elements of fast processing time, accurate results and less occupied resources. In particular, accurate results are always top priorities. A discovery motif has a fast runtime and takes up little resources, but inaccurate results are not appreciated compared to another discovery motif that has more accurate results even if it takes a lot of runtime and takes up more resources.

In 2002, Lin J, Keogh and colleagues proposed a solution to determine the effectiveness of a discovery motif. This method is built based on the determination of the efficiency constant as follows:

$$\text{Efficiency} = \frac{\boldsymbol{Number\ of\ Time\ call\ Eclidean\ dist}}{\boldsymbol{Number\ of\ Times\ brute - Force\ call\ Eclidean\ dist}}$$

In this section, we implement two Brute force and MK algorithms to discover motif in which two Euclid and DTW measurement are used to calculate the distance between the two strings. On this result, we will evaluate the accuracy and runtime of algorithms on real data sets.

3.1 Discovery 1-Motif with Brute force algorithm

Table 1: The Brute force algorithm

```
Algorithm    Brute Force Motif Discovery
Procedure    [$L_1$,$L_2$] = BruteForce_Motif(D)
In:       $D$: Database of Time Series
Out:      $L_1$,$L_2$: Locations for a Motif
1.        best-so-far = INF
2.        for $i$ = 1 to $m$
3.           for $j$ = $i$+1 to $m$
4.             if $d(D_i,D_j)$< best-so-far
5.               best-so-far = $d(D_i, D_j)$
6.               $L_1$ = $i$, $L_2$ =$j$
```

The brute force algorithm as shown in Table 1, whose runtime is $O(m^2)$ with m is the number of subsequences in the time series data. It is simply two nested loops that check sequentially every possible combination of other time series and give $\{L_1, L_2\}$ pairs the minimum distance between them. This algorithm results in a 1-motif.

In line 4 of Table 1, to calculate the distance $d\ (D_i, D_j)$, it is possible to use Euclid or DTW measurement and in line 6 for motif pairs (L_1, L_2) has the smallest distance between each other.

Brute force algorithm gives accurate results, but with increasing input data, runtime also increases. However, this algorithm is often used as a basis to evaluate the accuracy of results compared to other algorithms.

3.2 Discovery 1-Motif with MK algorithm

The highlight of the MK Algorithm [7] is to use multiple reference time series in the data set and perform distance calculations from these reference strings to all subsequences in the data set using standard deviations to sort distance of strings in the data set. The goal is to end the calculation and search process early, which reduces runtime. MK algorithm has high efficiency both in terms of accuracy and discovery motif time as shown in Table 2.

Table 2: The MK algorithm

```
Algorithm MK Motif Discovery
Procedure [$L_1$,$L_2$]= MK_Motif ($D$,$R$);$D$: Database of Time Series
best-so-far = INF
for $i$=1 to $R$
  $ref_i$ = a randomly chosen time series $D_r$ from $D$
  for $j$= 1 to $m$
    $Dist_{i,j}$ = $d(ref_i,D_j)$
    if $Dist_{i,j}$ < best-so-far
      best-so-far = $Dist_{i,j}$, $L_1=r$, $L_2=j$
      $S_i$= standard_deviation($Dist_i$)
find an ordering $Z$ of the indices to the reference time series in ref
 such that $S_{Z(i)} \geq S_{Z(i+1)}$
find an ordering  $I$  of the indices to the time series in $D$ such that
$Dist_{Z(1),I(j)} \leq Dist_{Z(1),I(j+1)}$
offset = 0, abandon = false
while abandon = false
   offset = offset + 1, abandon = true
   for $j$=1 to $m$
      reject = false
      for $i$=1 to $R$
         lower_bound = | $Dist_{Z(i),I(j)} - Dist_{Z(i),I(j+offset)}$ |
         if lower_bound > best-so-far
            reject = true, break
         else if  $i$ = 1
            abandon = false
            if reject = false
               if $d(D_{I(j)},D_{I(j+offset)})$< best-so-far
                  best-so-far = $d(D_{I(j)}, D_{I(j+offset)})$
                  $L_1=I(j)$,$L_2$= $I(j+offset)$
```

3.3 Our implementation use Euclidean and DTW on Brute force and MK algorithms

We experimented on two Brute-Force and MK algorithms for discovery motifs in which used Euclid and DTW measurements on real data sets and dimensional reduction data sets. Figure 4 illustrates the experimental model of algorithms.

Model parameters include:

- Size of sliding window *w* (number of points for a subsequence)
- Value *r*: Warping window used in Sakoe Chiba [15] and LB_Keogh limit technique.
- Value *R*: Number of reference strings used in MK algorithm
- PAA value: The number of points collected decreases into one point.
- SAX value: Number of break points according to Gauss distribution.

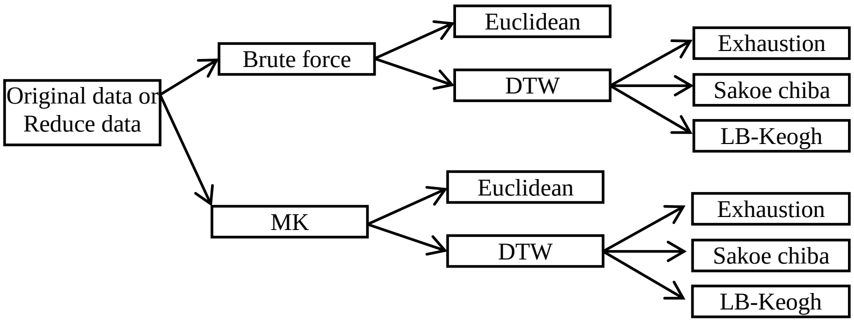


Figure 4: The tree diagram of the experimental model on the original data or reduced data by PAA method.

4 EXPERIMENTAL EVALUATION

We implemented the motif discovery algorithms: Brute-force and MK with Microsolft Visual C # and conducted the experiments on an Intel$^{®}$ Core ^{TM} i2 CPU T5870, 2GHz, RAM 4GB, Window 7.

In this experiment, we compared and evaluated when using Euclidean and DTW measurements for discovery motifs on the time series. In addition, we also use limited Sakeo Chiba and LB_Keogh techniques in the warping window and experiment on two data sets: EEG, Chromosome.

4.1 Experiment on EEG Data set

The original EEG dataset length is changed from: 500, 1000, the motif length is changed from: 80, 128 and warping window varies from 1 to 10.

We tested Brute force and MK algorithms and used Euclid and DTW measurements. Similarly we also tested on reduced dimension data with the same data set length and motif length as in the original data. The results are as shown in Table 3 and Figure 5.

Table 3: Experiment results on Brute force and MK algorithm when using Euclid and DTW measurements

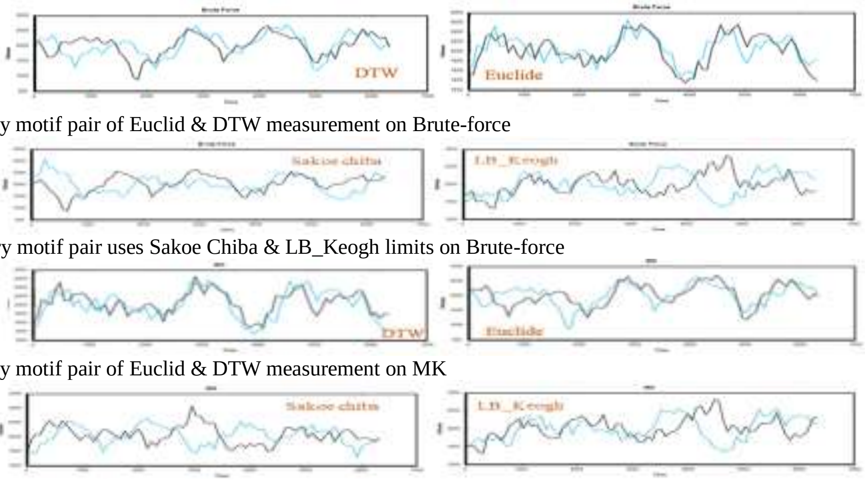
Algorithm	Measure	Bounding	Ref	R	Efficiency	Compare	Runtime (s)	BSF
Brute-force	Euclid	x	x	x	1	8410	0.0312	6.302501
	DTW	Exhaustion	x	x	1	8410	19.6729	4.920061
		Sakoe Chiba	x	1	1	8410	3.1328	0.070711
		LB_Keogh	x	1	1	8410	14.0781	12.4157
MK	Euclid	x	6	x	0.88079	77871	0.0469	6.302501
	DTW	Exhaustion	6	x	0.93815	82942	19.3145	4.920061
		Sakoe Chiba	6	1	0.02336	2065	0.2031	0.089443
		LB_Keogh	6	5	0.97233	85964	**63.648**	13.21055

In Table 3: Ref is the reference string number, used for MK algorithm. Here we take the Ref = 6 value as the reference string value (This value has been tested at [16]), R is the size of the warping window used in the Sakoe Chiba and LB_Keogh limits. BSF is the resulting motif.

Experimental results show that: on two Brute-force and MK algorithms, the cost of DTW measurement is higher than the Euclidean measurement. However, the resulting motif of the DTW measurement is better than the Euclidean measurement. The use of two techniques that limit Sakoe Chiba and LB_Keogh to DTW measurements on two algorithms to reduce computation time has been effective with R = 1. However, the algorithm MK limits LB_Keogh to a high cost. is 63,648 seconds. Discovery motif on MK algorithm has fast processing time and low resource utilization, has better efficiency than Brute-force algorithm (Efficiency <1). Similarly with data length 1000 points, MK algorithm gives better effect of Brute-force algorithm when using both Euclid and DTW measures.

Details of discovery motif on Brute-force and MK with data length: 500 points, motif length: 80 as shown in Figure 5.

Figure 5: The discovery motif pair of Euclid and DTW measurement on Brute-force and MK



(a) The discovery motif pair of Euclid & DTW measurement on Brute-force

(b) The discovery motif pair uses Sakoe Chiba & LB_Keogh limits on Brute-force

(c) The discovery motif pair of Euclid & DTW measurement on MK

(d) The discovery motif pair uses Sakoe Chiba & LB_Keogh limits on MK

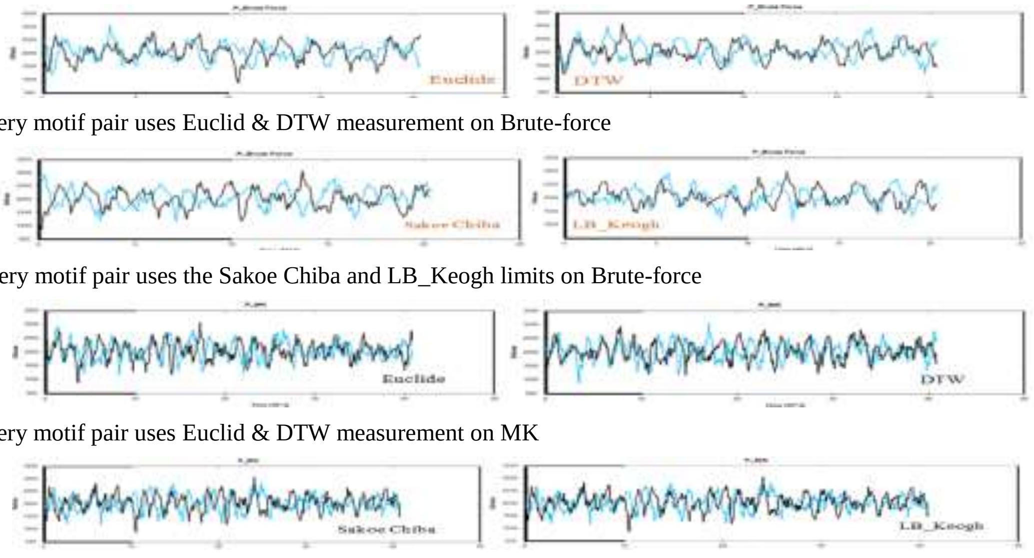
On reduced data EEG, there are results such as Table 4 and Figure 6

Table 4: Data length: 800 points, motif length 256, PAA: 32,64,128, measuring Euclid and DTW

Algorithm	Measure		PAA	SAX	Ref	R	Efficiency	Compare	Runtime (s)	BSF
Brute-force	Euclid		32	3	x	x	1	148240	0.1406	2.432447
			64	3	x	x	1	148240	0.1406	1.72
			128	3	x	x	1	148240	0.1523	1.216224
	DTW	Exhaustion	32	3	x	x	1	148240	**5.9258**	**0.466708**
			64	3	x	x	1	148240	23.2812	0.508959
			128	3	x	x	1	148240	88.965	0.488617
		Sakoe Chiba	32	3	x	5	1	148240	**4.1797**	0.138265
			64	3	x	5	1	148240	9.0938	0.08712
			128	3	x	5	1	148240	17.1406	0.053655
		LB_Keogh	32	3	x	5	1	148240	**1.1914**	0.463909
			64	3	x	5	1	148240	0.3281	0.093488
			128	3	x	5	1	148240	0.1094	0.001414
MK	Euclid		32	3	6	x	0.76096	112804	0.0625	4.049181
			64	3	6	x	0.36582	54229	0.0312	2.980448
			128	3	6	x	0.74391	110277	0.0781	4.312217
	DTW	Exhaustion	32	3	6	x	0.0191	2832	**1.4688**	**0.029514**
			64	3	6	x	0.01237	1834	3.8438	0.022273
			128	3	6	x	0.006	890	6.5156	0.019466
		Sakoe Chiba	32	3	6	5	0.98045	145342	6.7578	0.835464
			64	3	6	5	0.9811	145438	14.4609	0.748665
			128	3	6	5	0.98118	145450	23.4609	0.609098
		LB_Keogh	32	3	6	5	0.0625	203	0.0625	0.028709
			64	3	6	5	0.00042	62	0.0156	0.043301
			128	3	6	5	0.00115	171	0.0938	0.033541

Experimental results on reduced data in Table 4 show: the time of using DTW measurement is higher than the Euclidean measurement on Brute-force and MK algorithms. The discovery motif result of DTW measurement is better than the Euclidean measurement. The technical use of limited Sakoe Chiba with R= 5 on Brute-force for runtime is faster than exhaust. But on MK, limit Sakoe Chiba to R = 5 for a higher time than the exhausted method. However, using LB_Keogh limit on both Brute-force and MK algorithms, runtime is fast. The efficiency of MK algorithm is better than Brute-force algorithm when using both Euclid and DTW measurements.

Details of discovery motif on Brute-force and MK algorithm with data length: 800 points, motif length: 256 as shown in Figure 6



(a) The discovery motif pair uses Euclid & DTW measurement on Brute-force

(b) The discovery motif pair uses the Sakoe Chiba and LB_Keogh limits on Brute-force

(c) The discovery motif pair uses Euclid & DTW measurement on MK

(d) The discovery motif pair uses the Sakoe Chiba and LB_Keogh limits on MK

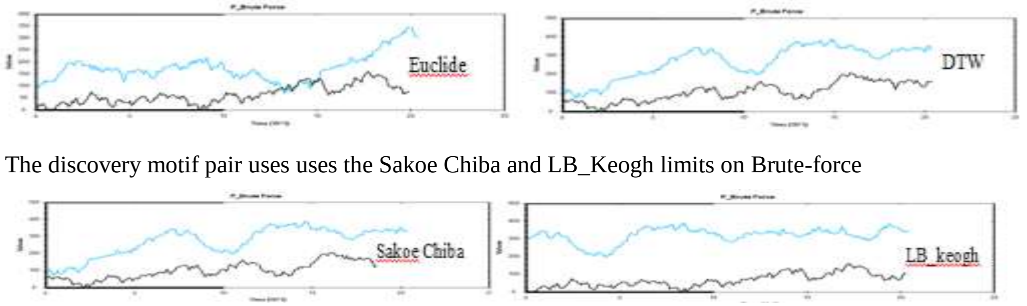
Figure 6: Experiment results on Brute-force and MK algorithms

4.2 Experiment on Chromosome Data set

On the Chromosome data set, we only experiment on the reduced data on both Brute-force and MK algorithms shown in Tables 5 and 6. The discovery motif details are shown in Figures 7 and 8.

Table 5: Data length: 800 points, motif length 256, measuring Euclid and DTW, Brute-force.

Brute-Force		PAA	SAX	Ref	R	Efficiency	Compare	Run time (s)	BSF
Euclid		32	3	x	x	1	148240	0.188	2.432447
		64	3	x	x	1	148240	0.2651	1.72
		128	3	x	x	1	148240	0.4058	1.216224
DTW	Exhaustion	32	3	x	x	1	148240	5.5889	0.362257
		64	3	x	x	1	148240	22.5967	0.357804
		128	3	x	x	1	148240	84.0680	0.328562
	Sakoe Chiba	32	3	x	5	1	148240	2.2812	0.090225
		64	3	x	5	1	148240	6.6797	0.039031
		128	3	x	5	1	148240	14.9297	0.032476
	LB_Keogh	32	3	x	5	1	148240	1.1562	0.12782
		64	3	x	5	1	148240	0.3281	0.015969
		128	3	x	5	1	148240	0.1094	0.002



(a) The discovery motif pair uses Euclid & DTW measurement on Brute-force

(b) The discovery motif pair uses uses the Sakoe Chiba and LB_Keogh limits on Brute-force

Figure 7: Experiment results on Brute-force

The parameters in Table 5 mean:

PAA: number of points reduced to one point, SAX: number of break points according to Gauss distribution, ref: number of strings as reference, used for MK algorithm. Here we take the Ref = 6 value as the reference string value (this value has been tested at [4]). R: Helical window width used in Sakoe Chiba and LB_Keogh limit techniques, Efficiency: algorithm performance index, Compare: number of comparisons. Runtime: the time the algorithm performs the search for the motif result (in seconds) and BSF: the result of the pair of motifs found.

The result of the pair of motifs obtained by Brute-force algorithm on Figure 7a when using both Euclide and DTW measurements, DTW for motif pairs of the same shape, results in BSF = 0.362257, while the result of the measurement Euclide has BSF = 2,432447.

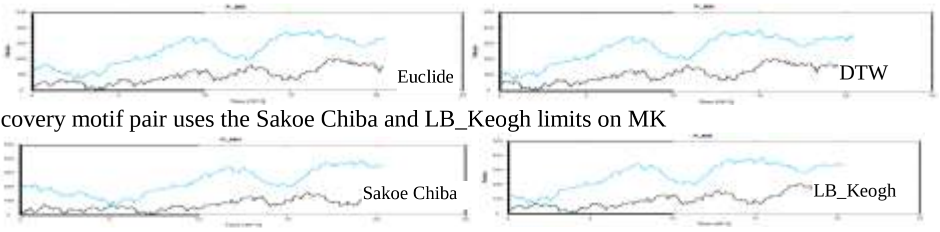
On Figure 7b, when using the Sakoe Chiba limit technique with R = 5, the resulting motif shape found to be similar to the exhausted DTW measurement. While using the LB_keogh limit technique with the limit of R = 5, the resulting pair of motifs found to be close to the Sakoe Chiba limit.

Table 6: Data length: 800 points, motif length 256, measuring Euclid and DTW, MK.

MK		PAA	SAX	Ref	R	Efficiency	Compare	Run time (s)	BSF
Euclid		32	3	6	x	0.47169	69924	0.0469	5.121227
		64	3	6	x	0.39577	58669	0.0469	5.438607
		128	3	6	x	0.35532	52672	0.0469	5.459352
DTW	Exhaustion	32	3	6	x	0.00157	233	1.4541	0.023133
		64	3	6	x	0.00126	187	3.5940	0.021651
		128	3	6	x	0.00023	34	6.1099	0.015436
	Sakoe Chiba	32	3	6	5	1.02498	151943	3.4531	0.535257
		64	3	6	5	0.97611	144699	8.2578	0.395917
		128	3	6	5	0.98046	145344	20.5312	0.390512
	LB_Keogh	32	3	6	5	0.0004	59	0.0605	0.059128

		64	3	6	5	0.00124	184	0.0312	0.040984
		128	3	6	5	0.00212	315	0.1094	0.013835

(a) The discovery motif pair uses Euclid and DTW measurement on MK



(b) The discovery motif pair uses the Sakoe Chiba and LB_Keogh limits on MK

Figure 8: Experiment results on MK

With the MK algorithm, the result of the pair of motifs was found in Figure 8a when both Euclide and DTW measurements were used, DTW for the pair of motifs of the same shape, BSF result = 0.023133, while the result of the Euclidean measurement then BSF = 5.121227.

In Figure 8b, using the limit technique of Sakoe Chiba with R = 5, the resulting motif shape found to be similar to the DTW exhausted and BSF = 0.535257. While using the LB_keogh limit technique with the limit of R = 5, the resulting pair of motifs are approximate to the exhausting DTW measurement.

We obtained the results when experimenting on the Chromosome dataset as follows:

• Time: calculation time and the resulting motif of DTW measurement higher than Euclidean measurement on both Brute-force and MK algorithms. Therefore, using Euclidean measurements on this dataset gives higher efficiency.

• Limited technique: the limited use of Sakoe Chiba technique with R = 9 gives the calculation time of the MK algorithm faster than Brute-force, but with LB_Keogh limit technique, the calculation cost is higher on two Brute-force and MK algorithms.

• Algorithm efficiency: MK algorithm always gives better effect of Brute-force algorithm on both measures of Euclide and DTW.

5 CONCLUSIONS

Through experimentation, we have some results as follows:

- Measurement: DTW measurement has a higher cost than Euclid measurement even though we use the Sakoe Chiba, LB_Keogh limiting techniques to speed up the calculation time, but the results found using DTW measurement are better than Euclidean measurements.
- Effectiveness: the MK algorithm is more effective than the Brute-force algorithm on the original data set and on the reduced dimension data set.
- Runtime: runtime of lower MK algorithm than Brute-force algorithm.
- LB_Keogh limit: with two algorithms MK and Brute-force using LB_Keogh limit technique, the motif found on MK algorithm has better results. However, if comparing the runtime between the Sakoe Chiba and LB_Keogh limits, the LB_Keogh limit is not highly effective on both algorithms.

Through experimenting, discovery problem motif on original data and dimensional reduction data on both algorithm MK and Brute-force. We found that the MK algorithm can be easy to find motifs in time series with better runtimes than Brute-force algorithms, especially when the data set is larger. Using DTW measurements often results in more accurate motifs than Euclidean measurements, but the runtimes must be accepted higher.

We continue to experimentally compare the following limits: LB Improved, FTW (Fast search method for dynamic Time Warping), and EDM [17]. Experiment on multiple data sets with different characteristics and sizes for more reliable conclusions.

REFERENCES

[1] C. Gruber, M. Coduro, B. Sick, "*Signature Verification with Dynamic RBF Networks and Time Series Motifs*," in Proc of 10th Int. Workshop on Frontiers in Handwriting Recognition, p. 2006.

[2] Y. Jiang, C. Li, J. Han, "*Stock temporal prediction based on time series motifs*," in Proc. of 8th Int. Conf. on Machine Learning and Cybernetics, 2009.

[3] K. Buza and L. S. Thieme, "*Motif-based Classification of Time Series with Bayesian Networks and SVMs*," in A. Fink et al. (eds.) Advances in Data Analysis, Data Handling and Business Intelligences, Studies in Classification, Data Analysis, Knowledge Organization. Springer-Verlag, 2010, pp. 105-114.

[4] A.Mueen, E.Keogh, Q.Zhu, S.Cash and B.West, "*Exact Discovery of Time Series Motifs*", SLAM International Conference on Data Mining (SDM09), 2009.

[5] P. Beaudoin, M. Panne, and S. Coros, "*Motion-Motif Graphs*", Symposium on Computer Animation 2008.

[6] T. Guyet, C. Garbay and M. Dojat, "*Knowledge construction from time series data using a collaborative exploration system*", Journal of Biomedical Informatics 40(6): 672-687 (2007).

[7] Eamonm Keogh, *Exact indexing of dynamic time warping*, in Knowledge and Information Systems (2004).

[8] J. Meng, J.Yuan, M. Hans and Y. Wu, "*Mining Motifs from Human Motion*", Proc. of EUROGRAPHICS, 2008.

[9] D. Minnen, C.L. Isbell, I. Essa, and T. Starner, "*Discovering Multivariate Motifs using Subsequence Density Estimation and Greedy Mixture Learning*", 22nd Conf. on Artificial Intelligence (AAAI'07), 2007.

[10] Keogh, E., & Pazzani, M. (2000) *Scaling up dynamic time warping for datamining applications*. In 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.285-289.

[11] Ghazi Al-Naymat, " *New methods for mining Sequential and time series Data*"
a novel approach to speed up dynamic time warping, 2009.

[12] E. Keogh, T. Palpanas, V.Zordan, D.Gunopulos, and M. Cardle, "*Indexing Large Human-Motion Databases*", Porceeding of the 30th International Conference on Very large Data Bases (VLDB04), 2004, pp.780-791.

[13] E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotra, "*Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases*", in Knowledge and Information System, vol.3, No.3, 2000, pp.263-286.

[14] E. Keogh, K.Charkrabarti, M.Pazzani, "*Locally adaptive dimensionality reduction for indexing large time series databases*", in Proc. of 2001 ACM SIGMOD Conference on Management of Data, 2001, pp.151-162.

[15] Sakoe H., Chiba S.(1978), *Dynamic programming algorithm optimization for spoken word recognition*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 26, 1, 43 – 49.

[16] J. Lin, E. Keogh, S. Lonardi, and P. Patel, *Finding motifs in time series*, Proc. of 2nd Workshop on Temporal Data Mining (KDD'02), 2002.

[17] D. T. Anh, N. V. Nhat, "*An Efficient Implementation of EDM Algorithm for Motif Discovery in Time Series data*", Int. J. Data, Modelling and Management, Vol. 8, No. 2, 2016.

SO SÁNH HIỆU QUẢ GIỮA HAI ĐỘ ĐO EUCLIDE VÀ DTW DÙNG TRONG KHÁM PHÁ MOTIF TRÊN CHUỖI THỜI GIAN

Tóm tắt. Nghiên cứu về cơ sở dữ liệu chuỗi thời gian dựa trên việc truy xuất hiệu quả các mẫu chưa biết và thường gặp trong chuỗi thời gian, được gọi là motif, đã thu hút nhiều sự chú ý của nhiều nhà nghiên cứu gần đây. Những motif này rất hữu ích cho việc khám phá dữ liệu và cho lời giải của nhiều bài toán trong các lĩnh vực ứng dụng khác nhau. Trong bài báo này, chúng tôi nghiên cứu và đánh giá hiệu quả của việc sử dụng cả hai phương pháp đo khoảng cách Euclide và Dynamic Time Warping (DTW), sử dụng thuật toán Brute-force và Mueen - Keogh (MK), trong đó thuật toán MK đã thực hiện hiệu quả về thời gian và độ chính xác của bài toán để khám phá các motif. Hiệu quả của phương pháp này đã được chứng minh thông qua thử nghiệm trên cơ sở dữ liệu thực.

Từ khóa. chuỗi dữ liệu thời gian, khám phá motif, độ đo DTW, độ đo Euclide, giới hạn Sakoe Chiba, giới hạn LB_Keogh.

Ngày nhận bài:17/01/2019

Ngày chấp nhận đăng:20/02/2019