**American Journal of Data Mining and Knowledge Discovery**
2019; 4(1): 32-45
http://www.sciencepublishinggroup.com/j/ajdmkd
doi: 10.11648/j.ajdmkd.20190401.16
ISSN: 2578-7810 (Print); ISSN: 2578-7837 (Online)

**Methodology Article**

# A New Similarity Measure for Time Series Data Mining Based on Longest Common Subsequence

**Gholamreza Soleimany, Masoud Abessi***

Department of Industrial Engineering, Yazd University, Yazd, Iran

**Email address:**
gholam_soleimani@yahoo.com (G. Soleimany), mabessi@gmail.com (M. Abessi)

*Corresponding author

**To cite this article:**
Gholamreza Soleimany, Masoud Abessi. A New Similarity Measure for Time Series Data Mining Based on Longest Common Subsequence. *American Journal of Data Mining and Knowledge Discovery*. Vol. 4, No. 1, 2019, pp. 32-45. doi: 10.11648/j.ajdmkd.20190401.16

**Received**: May 3, 2019; **Accepted**: June 3, 2019; **Published**: June 20, 2019

**Abstract:** In this research, a new similarity measurement method that named Developed Longest Common Subsequence (DLCSS) is suggested for time series data mining. The main idea of the DLCSS is using the logic of the Longest Common Subsequence (LCSS) method and the concept of similarity in time series data. In most studies related to time series data mining, referred to the LCSS and Dynamic Time Warping (DTW) methods as the best and most usable for similarity measurement methods, but the LCSS is intrinsically designed to measure the similarity of two sequences of character, which later was developed for time series by defining and determining the similarity threshold. The value of similarity threshold has huge impact on the quality of time series data mining. In the DLCSS by defining two similarity thresholds and determining the values of them, this defect is eliminated. The performance of the DLCSS will be compared with the LCSS and DTW in time series data mining by the Query by content and K-medoids Clustering techniques on 23 datasets from the UCR datasets. The result shows that it is possible to claim that the performance of the DLCSS is better than the LCSS and DTW with 90% confidence.

**Keywords:** Time Series, Data Mining, Similarity Measurement, Longest Common Subsequence, Dynamic Time Warping, Developed Longest Common Subsequence

## 1. Introduction

Time series data is a set of ordered numbers that expresses the temporal properties of the objects at any moment of time [1]. Time series data almost exist in all areas, as an example in the medical field such as the heart rate data, the intensity breathing data and the neurotoxicity of the brain for a period of time, in climate field such as the daily temperature data of a location and the daily humidity of a location, in the sales field such as daily, weekly, monthly or annual sales and in other different fields. Time series data have three important features. The 1st ones is to have a high dimension, so that sometimes a time series data can be have hundreds or more member and this occupies high memory space and reduces the speed of computing time series data mining. The 2nd ones is data-dependency, so that this feature plays a significant role in mining time series. Because the value of each member of a time series is influenced by the value of its former members, so it should be needed to carefully determine appropriate mathematical and statistical relationships. The 3rd ones is the need for their constant continuation update in most real applications [2-5].

Data mining is a particular importance way for discovering knowledge from a wealth of data, so that the use of various data mining techniques such as Classification, Clustering, Rule deduction, the Query by content, Forecasting in the different fields like production, medicine, social, meteorology, stock exchange, sales, customer service and etc. are increasing [2].

Time series data mining process is hard and special, because data mining techniques are specially designed for fixed data, and it needs to make changes to the corresponding algorithms for time series data mining [6]. These changes are reducing dimension of time series and choosing appropriate similarity measurement method. The dimension reducing of a time series means indexing. It's aim is reduction of calculation time and it should be done in such a way that the amount of lost knowledge due to the reduction of the time series do not deviate from

achieving the right result [7]. A survey of various types of indexing methods has been carried out by Aghabozorgi et al. (2015) [8]. The choosing appropriate similarity measurement method is determining the appropriate time series similarity measurement method for Time series data mining, which is important effective factor in the quality of results. It should be noted that provide a suitable method for measuring the similarity of time series is one of the issues that has been widespread in time series data mining research in recent years [8]. Furthermore Aghabozorgi et al. (2015) showed that the Longest Common Subsequence (LCSS) and Dynamic Time Warping (DTW) methods have been used in more research and have a much better performance than other methods [8].

So with these descriptions and due to the importance and impact of the similarity measurement method in time series data mining, in this research a new method for time series similarity measurement is proposed and the performance of this method compared with the performance of the LCSS and the DTW similarity measurement methods.

In the following, at first the concepts of similarity, kinds of similarity measurement methods and the relations to calculate some of them especially the LCSS and the DTW methods discuss. After that the developed LCSS-based methods and their specifications are described. Then the proposed method for measuring similarity of time series is presented. After that, the Query by content and K-medoids techniques are done with the proposed, LCSS and DTW methods for the time series datasets and final the results are analyzed.

# 2. Similarity and History of Similarity Measurement Methods

As before discussed, one of the important problems in time series data mining is the similarity problem. Based on the research, Similarity in time series is defined as point-to-point similarity and flexible similarity (one point to several points or several points to one point). There is another definition for similarity that define as similarity in time, similarity in shape and similarity in model. Similarity in time means that the similarity between two time series based on similarity at any given moment in time. Similarity in shape is the similarity between two time series based on the similarities between the following subsequence and the similarity in model also means the uniformity of the parameters and the uniformity of the fitted model to two time series [9].

On the other hand, there are generally two approaches for time series similarity measurement, the Whole matching approach and the subsequence matching approach. In the Whole matching approach, total length of time series are used, that is if the length of each time series is equal to m, all m data of the first time series and all m data of the second time series are used. In the subsequence matching approach, the time series have different lengths and similarity measurement between them is based on the similarity between the following subsequences. If the length of them are n and m respectively, and n <m then subsequences with the Consecutive data of length n from the time series with greater length will select and the similarity of each of these subsequences with smaller time series is measured. The most similarity obtained is considered as the similarity of the two time series [10].

The similarity measures can be also categorized into four categories: 1. Shape based distance measure, 2. Edit based distance measure, 3. Feature based distance measure and 4. model based distance measure [11].

In the following some of the famous distance measuring methods of time series in the domain of shape-based, edit-based and feature-based distance measure are presented and the strengths and weaknesses of them will be expressed.

### *2.1. Shape Based Distance Measures Group*

This Group of measures is based on directly use the raw values and shapes of the time series in different manners. Below, the most commonly used methods of this group are discussed. Suppose that $TS_X = (x_1, x_2, \dots, x_n)$ and $TS_Y = (y_1, y_2, \dots, y_n)$ represent the time series X and Y, respectively with length n.

#### *2.1.1. Distance Measurement Method Based on Lp-Norms*

One of the most well-known shape based distance measurement that had been used in investigations related to time series data mining is the Lp-Norms method, which is considered as a strict metric method, only use for time series with equal lengths and it is point-to-point similarity type [12]. In this method the distance between $TS_X$ and $TS_Y$ is calculated by relation (1).

$$LP(TS_X, TS_Y) = \left(\sum_{i=1}^{n}(x_i - y_j)^p\right)^{\frac{1}{p}} = ((x_1 - y_1)^p + (x_2 - y_2)^p + \cdots + (x_n - y_n)^p)^{1/p} \quad (1)$$

In this relation, p is a natural number and when p=1 it is called the coordinate relation (Manhattan relation) and when p=2 it is known as the Euclidean relation. While two time series are similar in shape and this similarity occurs with a time delay then this relationship can not identify this similarity and it is the main weakness of this method.

#### *2.1.2. Short-time Series Method*

In Short Time Series or STS method, each time series is considered as a linear function. In this method the distance between $TS_X$ and $TS_Y$ is calculated by relation (2), so that the parameter $t_i$ represent the time of the measurement of the $i^{th}$ data [13]. Weakness of STS method is same as the weakness of Lp-Norm method.

$$d_{STS}(TS_X, TS_Y) = \sqrt{\sum_{i=1}^{n}\left(\frac{y_{(i+1)} - y_{(i)}}{t_{(i+1)} - t_i} - \frac{x_{(i+1)} - x_{(i)}}{t_{(i+1)} - t_i}\right)} \quad (2)$$

#### *2.1.3. Dynamic Time Warping Method (DTW)*

The DTW method is a method that has been able to overcome the weakness of the above methods [14]. Because sometimes there is time series that are roughly same in general but this

similarity does not coincide along the axis of time. In fact, the DTW method is presented to calculate the similarity between two time series with different lengths and has a significant difference with the previous methods. This difference is the possibility of lengthening the length of a time series by dragging it (repeating some of its data which is similar to the other time series data). This method uses a backward relation to calculate the non-similarity between two time series with lengths n and m respectively, as $\mathrm{DTW}(\mathrm{TS_x}, \mathrm{TS_y}) = \mathrm{M}(\mathrm{n}, \mathrm{m})$ which that M(n,m) calculated by the relation (3).

$$\mathrm{M}(\mathrm{i},\mathrm{j}) = \begin{cases} (\mathrm{x_i} - \mathrm{y_j})^2 ; \mathrm{i} = 1, \mathrm{j} = 1 \\ (\mathrm{x_i} - \mathrm{y_j})^2 + \mathrm{M}(\mathrm{i}, \mathrm{j}-1);\ \mathrm{i} = 1, \mathrm{j} \geq 2 \\ (\mathrm{x_i} - \mathrm{y_j})^2 + \mathrm{M}(\mathrm{i}-1, \mathrm{j});\ \mathrm{i} \geq 2, \mathrm{j} = 1 \\ (\mathrm{x_i} - \mathrm{y_j})^2 + \mathrm{Min}\begin{cases} \mathrm{M}(\mathrm{i}-1,\mathrm{j}) \\ \mathrm{M}(\mathrm{i},\mathrm{j}-1) \\ \mathrm{M}(\mathrm{i}-1,\mathrm{j}-1) \end{cases} ; \mathrm{i} \geq 2, \mathrm{j} \geq 2 \end{cases} \quad (3)$$

With this explanations, the result of the DTW method is $\mathrm{DTW}(\mathrm{TS_x}, \mathrm{TS_y})$ and a sequence with paired elements and the length r, where each paired element represents the data of first and second time series respectively, that are same (very close together) and the length of this sequence is certainly greater than or equal to $\mathrm{Max}(\mathrm{n}, \mathrm{m})$.

In order to comparable the DTW of two time series with the DTW of two other time series, the relation $\mathrm{dissim}(\mathrm{TS_x}, \mathrm{TS_y}) = \sqrt{\frac{\mathrm{DTW}(\mathrm{TS_x}, \mathrm{TS_y})}{||\mathrm{r}||}}$ is used.

In general, this method has a better performance than other time series measurement methods and has wider application [8].

### *2.2. Edit Based Measurement Method Group*

The edit based measurement methods group was originally presented to calculate the similarity between two sequences of characters, and based on the count of the minimum number of editing operations necessary (including removal, placement, and insertion) to convert a sequence to another sequence. In the following some of the most usual methods of this group are discussed. To continue suppose that $\mathrm{S_X} = (\mathrm{x_1}, \mathrm{x_2}, \ldots, \mathrm{x_n})$ and $\mathrm{S_Y} = (\mathrm{y_1}, \mathrm{y_2}, \ldots, \mathrm{y_m})$ are two sequences of characters.

#### *2.2.1. Levenshtien Distance Measurement Method*

The Levenshtine distance measurement method was presented by a Russian scientist Vladimir Levenshtine and it is widely used in spelling, speech recognition, DNA analysis, and plagiarism detection [15]. While the length of two sequences are n and m respectively, then $\mathrm{Lev}(\mathrm{S_x}, \mathrm{S_y}) = \mathrm{M}(\mathrm{n}, \mathrm{m})$ and $\mathrm{M}(\mathrm{n}, \mathrm{m})$ is calculated from the relation (4), so that $\mathrm{Sim}(\mathrm{x_i}, \mathrm{y_j}) = \begin{cases} 0 & ; \ \mathrm{x_i} = \mathrm{y_j} \\ 1 & ; \ \mathrm{x_i} \neq \mathrm{y_j} \end{cases}$.

$$\mathrm{M}(\mathrm{i},\mathrm{j}) = \begin{cases} \mathrm{i};\ \mathrm{if}\ \mathrm{j} = 0 \\ \mathrm{j};\ \mathrm{if}\ \mathrm{i} = 0 \\ \mathrm{min}\begin{cases} \mathrm{M}(\mathrm{i}-1,\mathrm{j}) + 1 \\ \mathrm{M}(\mathrm{i},\mathrm{j}-1) + 1 \\ \mathrm{M}(\mathrm{i}-1,\mathrm{j}-1) + \mathrm{Sim}(\mathrm{x_i}, \mathrm{y_j}) \end{cases} ; \ \mathrm{Otherwise} \end{cases} \quad (4)$$

This method is inherently created to compare two sequences of characters but it can be used for two time series by define the similarity threshold. This method is rarely used in time series data mining.

#### *2.2.2. Longest Common Subsequence Method (LCSS)*

The LCSS method is a classic problem in computer science. The task is to find the longest common subsequence of two sequences. The most important feature of this method is that it can be ignore noise and distortion values. This method is inherently created to compare two sequences of characters. The similarity in this method defines as the same of two characters of two sequences and $\mathrm{LCSS}(\mathrm{S_x}, \mathrm{S_y}) = \mathrm{M}(\mathrm{n}, \mathrm{m})$, So that M(n,m) is calculated by the relation (5) and $0 \leq \mathrm{M}(\mathrm{n}, \mathrm{m}) \leq \min(\mathrm{n}, \mathrm{m})$.

$$\mathrm{M}(\mathrm{i},\mathrm{j}) = \begin{cases} 0;\ \mathrm{i} = 0.\mathrm{or}.\mathrm{j} = 0 \\ 1 + \mathrm{M}(\mathrm{i}-1,\mathrm{j}-1); \mathrm{x_i} = \mathrm{y_j}, \mathrm{i} \geq 1.\mathrm{or}.\mathrm{j} \geq 1 \\ \mathrm{Max}\begin{cases} \mathrm{M}(\mathrm{i}-1,\mathrm{j}) \\ \mathrm{M}(\mathrm{i},\mathrm{j}-1) \end{cases} ;\ \mathrm{x_i} \neq \mathrm{y_j}, \mathrm{i} \geq 1.\mathrm{or}.\mathrm{j} \geq 1 \end{cases} \quad (5)$$

The relation $\mathrm{Sim}(\mathrm{S_x}, \mathrm{S_y}) = \frac{2*\mathrm{LCSS}}{\mathrm{m}+\mathrm{n}}$ is used to comparable the LCSS of two time series with the LCSS of two other time series, which is within the range 0 to 1. The closer to one, the two sequences are more similar.

In order to use the LCSS method for numerical sequences (time series), changes have been made in how to determine the similarity of the two data. So when the absolute value of the difference between the two data of two time series is less than or equal the similarity threshold then it is considered to be similar, otherwise the two data are not similar [16-17]. With this description, the relation (5) is rewritten as relation (6).

$$\mathrm{M}(\mathrm{i},\mathrm{j}) = \begin{cases} 0;\ \mathrm{i} = 0.\mathrm{or}.\mathrm{j} = 0 \\ 1 + \mathrm{M}(\mathrm{i}-1,\mathrm{j}-1); |\mathrm{x_i} - \mathrm{y_j}| \leq \in, \mathrm{i} \geq 1.\mathrm{or}.\mathrm{j} \geq 1 \\ \mathrm{Max}\begin{cases} \mathrm{M}(\mathrm{i}-1,\mathrm{j}) \\ \mathrm{M}(\mathrm{i},\mathrm{j}-1) \end{cases} ; \in < |\mathrm{x_i} - \mathrm{y_j}|, \mathrm{i} \geq 1.\mathrm{or}.\mathrm{j} \geq 1 \end{cases} \quad (6)$$

The similarity threshold is $\in$. The logic used in this relation can be displayed in the Figure 1.

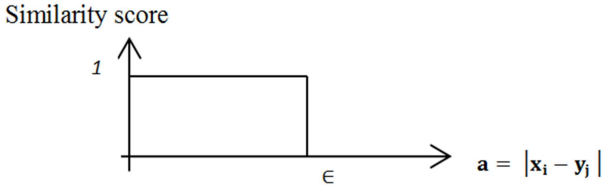


***Figure 1.*** *Conceptual description of similarity threshold in LCSS*

The result of LCSS is influenced by the value of $\in$, such that smaller value of $\in$, the smaller LCSS, and larger value of $\in$, the larger LCSS. The appropriate value of $\in$ depends on the nature of the data, but in the absence of any knowledge of the dataset and its features, the use of this method practically hasn't any conceptual.

#### *2.2.3. Edit Distance for Real Sequence Method (EDR)*

In this method, the identical of characters of two sequences

are the criterion to calculate the number of changes that needed to same two sequences to each other.

As defined $EDR(S_X, S_Y) = M(n, m)$ and M (n, m) is calculated from the relation (7) and $SC = \begin{cases} 0; & x_i = y_j \\ 1; & x_i \neq y_j \end{cases}$.

$$M(i,j) = \begin{cases} i; \text{ if } j = 1 \\ j; \text{ if } i = 1 \\ \text{Min} \begin{cases} M(i-1, j-1) + SC \\ M(i-1, j) + 1 \quad ; \text{ otherwise} \\ M(i, j-1) + 1 \end{cases} \end{cases} \quad (7)$$

By definition of the similarity threshold, this method can be used to measure the time series distance, but this method has limited used in time series data mining [17].

#### *2.2.4. Edit Distance with Real Penalty Method (ERP)*

The ERP method is the adaptation to the edit distance which is combination of the DTW and the EDR [18]. It used to measure the distance of time series with unequal lengths. In this method $ERP(TS_X, TS_Y) = M(n, m)$ and M(n, m) is calculated from the relation (8):

$$M(i,j) = \begin{cases} \sum_{k=1}^{i} |(x_k - g| \; ; \text{ if } j = 0 \\ \sum_{k=1}^{j} |(y_k - g| \; ; \text{ if } i = 0 \\ \text{Min} \begin{cases} M(i-1, j-1) + |x_1 - y_1| \\ M(i-1, j) + |x_1 - g| \quad ; \text{ otherwise} \\ M(i, j-1) + |y_1 - g| \end{cases} \end{cases} \quad (8)$$

In the above relation, g is a constant value that represents the amount of fines and is determined by the user. This method has limited used in time series data mining.

### *2.3. Feature Based Distance Measurement Group*

The feature-based distance measures focus on extracting a set of features from time series and calculating the similarities between these features, rather than using the raw data of those time series. Suppose that $TS_X = (x_1, x_2, \ldots, x_n)$ and $TS_Y = (y_1, y_2, \ldots, y_n)$ represent the time series X and Y respectively.

#### *2.3.1. Pearson Correlation Coefficient and Related Coefficient*

Pearson correlation coefficient is one of the feature based distance methods and uses the relation (9). In this relation $\mu_x$ and $\mu_y$ represent the average of the first and second time series data respectively, $Sd_x$ and $Sd_y$ represent the standard deviation of the first and second time series data respectively.

$$PCC(TS_X, TS_Y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{Sd_x \times Sd_y} \quad (9)$$

Based on this relation, two distance metrics were defined which are $d_{PC1}(TS_X, TS_Y) = (\frac{1-PCC}{1+PCC})^{\beta}$ and $d_{PC2}(TS_X, TS_Y) = 2(1 - PCC)$, so that the value of β is defined by the user.

Note that the length of two time series must be equal in these relations.

#### *2.3.2. Cosine Angle*

The root of $d_{PC2}(TS_X, TS_Y)$ is called Cosine Angle and calculated by $CA(TS_x, TS_y) = \sqrt{2(1 - \frac{\sum_{i=1}^{n} x_i \times y_j - n\, \mu_x \mu_y}{n \times Sd_x \times Sd_y})}$.

Note that the length of two time series must be also equal in this relation. The weakness of these methods is like the LP-Norm method.

Interestingly, the performance of all above methods is such that it can not be specifically stated that a particular method is appropriate for any time series databases. In other words, based on the research carried out, it can be concluded that each one is good for a group of data set and is not good for the rest of the data set and it showed that the DTW and the LCSS methods are widespread used and they have better performance than other methods [8, 18-26].

So the purpose of this research is to develop the LCSS method to measure the similarity of time series. So before propose the new method, refer to all LCSS-based methods.

## 3. LCSS-based Methods

### *3.1. Constrained Longest Common Subsequence Metho*

The Constrained Longest Common Subsequence (C-LCSS) method is a method that calculates the Longest common subsequence of two sequences in relation to a 3rd sequence. As defined while $S_X$ and $S_Y$ are two input sequences and B is a finite sequence with length r, then the constrained longest common subsequence is a subsequence of the two input sequences and including B which has the longest length. The C-LCSS method has limited used in the consistency of two biological sequences with a common and assumed structure. The C-LCSS method does not use as a measure of distance in time series data mining [27].

### *3.2. Multiple Longest Common Subsequence Method*

The Multiple Longest Common Subsequence (MLCSS) method is a method that calculates the longest common subsequence of more than two sequences. As defined, while $S_1$, $S_2$, ,..., $S_k$ denote the K input sequence so that k >2, this method try to find the longest common subsequence of these sequences. This method is considered as a Np-Hard problem for k> 3, and it is necessary to use heuristic methods to solve it. Meanwhile, this method doesn't use in time series data mining [28-29].

### *3.3. Multiple Longest Common Subsequence Method*

The Weighted Longest Common Subsequence (WLCSS) or the Heaviest Common Subsequence (HCSS) is a method that calculates the longest common subsequence of two sequences with highest weight. In this method, each character has a positive weight and the purpose is to determine the common subsequence of two sequences so that this subsequence has the maximum weight of all the available subsequences. Due to the nature of this method, it can not be used in time series data mining [30].

### *3.4. Flexible Longest Common Subsequence Method*

The Flexible Longest Common Subsequence (FLCSS) is a

new type of longest common subsequence method that seeks to find the common subsequence of two sequences with highest consequence points. In other words, when sequencing is important this method can be used. But, the arrangement of the common subsequence is not important in time series data mining, so this method is practically not used in time series data mining [31].

### *3.5. Longest Common Subsequence with Gapped Constraint Method*

The Longest Common Subsequence with Gapped Constraint (LCSSGC) method is a modified method of LCSS. while A and B are two input sequences and C is a restriction sequence with a gap list so that the lengths of these sequences are m, n and r, respectively, the LCSSGC problem is to find the longest subsequence such as Z of the sequences A , B and C [32]. Due to the nature of this method, it can not be used in time series data mining.

In a general summary of all developed methods based on the LCSS method, they can't be used in time series data mining like the CLCSS, the WLCSS, the FLCSS, and the LCSSGS methods, or they use only to determine the representation of several time series like the MLCSS method.

## 4. Proposed Method for Measuring the Similarity of Time Series

As will be shown in section 6.1, the sensitivity of LCSS method to the similarity threshold is very high so in this research in order to reduce this sensitivity and increase the quality of the results of data mining processes such as the Query by content and clustering techniques, a new method is proposed which is based on the LCCS'logic and is named "Developed Longest common Subsequence" or "DLCSS".

The DLCSS method uses two similarity thresholds, the first similarity threshold $\in_1$ is used to recognize the definite similarity of two data and the second similarity threshold $\in_2$ is used to detect the conditional similarity of the two data. Some conditions must be met for each of these cases. The relation (10) shows how to calculate the DLCSS.

$$M(i,j)=\begin{cases}0 & ; i=0.\,or.\,j=0\\ 1+M(i-1,j-1); |x_i-y_j|\le \in_1, i,j\ge 1\\ \mathrm{Max}\begin{cases}\frac{\in_2-a}{\in_2-\in_1}+M(i-1,j-1)\\ M(i-1,j)\\ M(i,j-1)\end{cases} & ; \in_1 < |x_i-y_j|\le \in_2, i,j\ge 1\\ \mathrm{Max}\begin{cases}M(i-1,j)\\ M(i,j-1)\end{cases}; \in_2 < |x_i-y_j|, i,j\ge 1\end{cases} \tag{10}$$

$$a=|x_i-y_j|$$

$$i=1,2,\dots,m, j=1,2,\dots,n$$

$$0 \ll DLCSS(TS_x,TS_y)=M(n,m)\le min(n,m)$$

To better understanding the DLCSS method, look at the Conceptual description of similarity threshold of the DLCSS in Figure 2.

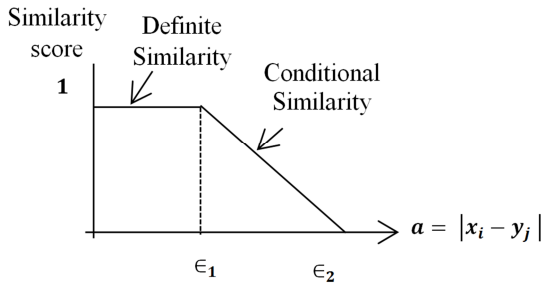


***Figure 2.*** *Conceptual description of similarity threshold in DLCSS.*

Unlike the LCSS that represents the length of the longest common subsequence and is a natural number between zero and min(m,n), the DLCSS represents the similarity score between two time series and can gives a real number between zero and min(m,n).

In contrast to the LCSS, the DLCSS does not have the rigid view (zero and one) to similarity, so while a data is a bit farther away but it is closer than the other adjacent data then it have chance to participate in similarity.

The logic used in DLCSS method is as follows:

a) The two data of two time series are certainly similar, if the absolute value of difference between these data is smaller or equal to $\in_1$. In this case, one unit will be added to the similarity score to the state of the preceding two data.

b) The two data are maybe similar, if the absolute value of the difference between these data is larger than $\in_1$ and smaller or equal to $\in_2$. This condition may be correct with respect to the status of the data before them. If this condition is correct, then the value that is added to similarity score is a fraction of one which is exactly equal to $\frac{\in_2-a}{\in_2-\in_1}$.

c) The two data of the two series are not definitely similar,

if the absolute value of difference between these data is greater than $\in_2$, then the similarity score is equal to the maximum similarity score before them.

# 5. Performance Evalution Approach

In this study, 23 series of time series data sets from the UCR data set are used, their name and specifications are presented in Table 1. Each time series data set has two distinct subsets, which are the training data set and the experimental data set. In each of the subsets, the class of each time series is specified. For example the "statistical control" data set has 6 clusters (class), and the length of each time series is 60 and the number of time series in the training data set and experimental data set are 300 and 300 respectively.

In this research, the performance of the LCSS and the DTW methods is compared to the proposed method on these data sets by the Query by content and K-medoids clustering techniques.

The Query by content technique has four steps in this research. First, the similarity of any time series of experimental dataset is measured by similarity measurement method with any time series of training dataset. Second, the most similar time series of training dataset to this time series is determined. Third, the class of time series is same as the class the most similar time series of training dataset. Four, the accuracy index is calculated. So that the accuracy is the ratio of the number of time series of experimental dataset that their class is correctly determined to the total number of time series of experimental dataset. This accuracy is the performance of the Query by content technique.

The K-medoids clustering technique in this research is used in two steps. In the first step, this technique run on the training data set and based on the accuracy of clustering, the best number of clusters and the representative of those clusters are selected. The accuracy clustering index in this process is the ratio of the number of time series of training data set that correctly assigned to the right cluster to the total number of time series of training dataset. In the second step, based on the best number of clusters and the cluster representative obtained from the first step, experimental data sets are grouped and the accuracy of these grouping is calculated as the ratio of the number of time series of experimental dataset that correctly assigned to the right cluster to the total number of time series of experimental dataset. These accuracy represent the performance of the K-medoids clustering technique.

All of the Programming that needed was written by MATLAB software.

***Table 1.*** *Name and specification of time series datasets that used in this research.*

| row | Database name | K | L | N1 | N2 |
|---|---|---|---|---|---|
| 1 | Statistical Control | 6 | 60 | 300 | 300 |
| 2 | GP | 2 | 150 | 50 | 150 |
| 3 | CBF | 3 | 128 | 30 | 900 |
| 4 | ECG | 2 | 96 | 100 | 100 |
| 5 | Face4 | 4 | 350 | 24 | 88 |
| 6 | Medical | 10 | 99 | 381 | 760 |
| 7 | Sweedian | 15 | 128 | 500 | 625 |
| 8 | OSU | 6 | 427 | 200 | 242 |
| 9 | Adiac | 37 | 176 | 390 | 391 |
| 10 | Beef | 5 | 470 | 30 | 30 |
| 11 | Lighting | 7 | 319 | 70 | 73 |
| 12 | Fish | 7 | 463 | 175 | 175 |
| 13 | 50words | 50 | 270 | 450 | 455 |
| 14 | Trace | 4 | 275 | 100 | 100 |
| 15 | Lighting7 | 7 | 319 | 70 | 73 |
| 16 | Distal | 7 | 80 | 139 | 400 |
| 17 | Italy power demand | 2 | 24 | 67 | 1029 |
| 18 | Middle-P-T | 7 | 80 | 154 | 399 |
| 19 | Plane | 7 | 144 | 105 | 105 |
| 20 | Car | 4 | 577 | 60 | 60 |
| 21 | Olive Oil | 4 | 570 | 30 | 30 |
| 22 | Diatom Size Reduction | 4 | 345 | 16 | 306 |
| 23 | Gun-Point | 2 | 150 | 50 | 150 |

K: Number of cluster.
L: length of Time series.
N1: Number of Time series in Training database.
N2: Number of Time series in Experimental database

# 6. Experimental Results

### *6.1. The Query by Content Results*

In this section, the results of the implementation of the Query by content technique by using the LCSS, DTW and DLCSS method as the similarity measurement are presented in Tables 2, 3, 4 respectively and the results are analyzed.

In Table 2, for example for statistical control dataset, the class of 97.33% of time series of the experimental data set as compared to the class of time series of training data set is correctly recognized. As you can see, the clustering accuracy of some datasets is very low, such as OSU dataset with 46.28% and Middle-P-T dataset with 58.4%. In addition, this method has been able to take 80.1% accuracy for all the datasets to

determining the correct class of the time series of experimental dataset.

***Table 2.*** *Accuracy of recognition of the correct Class of the test series of experimental dataset by implementation of the Query by content technique with the DTW method.*

| Row | database name | Accuracy% | Row | database name | Accuracy% |
|---|---|---|---|---|---|
| 1 | statistical control | 97.33 | 13 | 50words | 66.39 |
| 2 | GP | 90.67 | 14 | Trace | 100 |
| 3 | CBF | 99 | 15 | Lighting7 | 68.49 |
| 4 | ECG | 79 | 16 | Distal | 70.5 |
| 5 | Face4 | 81.82 | 17 | Italy power demand | 93.97 |
| 6 | Medical | 65.39 | 18 | Middle-P-T | 58.4 |
| 7 | Sweedian | 72.16 | 19 | Plane | 100 |
| 8 | OSU | 46.28 | 20 | Car | 71.67 |
| 9 | Adiac | 74.03 | 21 | Olive Oil | 83.33 |
| 10 | Beef | 63.33 | 22 | Diatom Size Reduction | 96.41 |
| 11 | Lighting | 68.49 | 23 | Gun-Point | 90.67 |
| 12 | Fish | 75.43 | Average accuracy% | | 80.1 |

***Table 3.*** *Accuracy of recognizing the correct Class of the time series of experimental dataset by implementation of the Query by content technique with the LCSS method and different values of ∈.*

| row | database name | ∈ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **0.05** | **0.1** | **0.15** | **0.20** | **0.25** | **0.30** | **0.35** |
| 1 | statistical control | 69.67 | 86.67 | 87 | 91.67 | 93 | 93 | 94.67 |
| 2 | Gun-Point | 92.67 | 98 | 98.67 | 97.33 | 92.67 | 88.67 | 85.33 |
| 3 | CBF | 96.89 | 98.11 | 99.56 | 99.44 | 99.67 | 99.78 | 99.89 |
| 4 | ECG | 77 | 85 | 86 | 87 | 90 | 91 | 91 |
| 5 | Face4 | 81.82 | 89.77 | 92.05 | 93.18 | 94.5 | 92.05 | 94.32 |
| 6 | Medical | 53.92 | 57.76 | 60.13 | 61.84 | 62.33 | 63.16 | 61.71 |
| 7 | Sweedian | 43.2 | 75.2 | 82.4 | 83.5 | 84.62 | 84.96 | 82.72 |
| 8 | OSU | 62.4 | 67.36 | 68.6 | 69.42 | 69.83 | 68.18 | 68.18 |
| 9 | Adiac | 77.92 | 66.88 | 55.19 | 48.86 | 42.7 | 33.38 | 26.23 |
| 10 | Beef | 56.67 | 70 | 70 | 76.67 | 73 | 63.33 | 56.67 |
| 11 | Lighting | 60.27 | 56.16 | 69.86 | 71.33 | 75.34 | 75.34 | 68.49 |
| 12 | Fish | 84 | 88 | 89.14 | 84.57 | 81.69 | 77.14 | 73.14 |
| 13 | 50words | 60.5 | 68.35 | 73.11 | 73.51 | 73.95 | 75.35 | 77.03 |
| 14 | Trace | 92 | 97 | 100 | 99 | 98 | 96 | 94 |
| 15 | Lighting7 | 60.27 | 56.16 | 69.86 | 71.23 | 75.34 | 75.34 | 68.49 |
| 16 | Distal | 69.5 | 73.5 | 74 | 75.25 | 75.5 | 76 | 76.25 |
| 17 | Italy power demand | 78.52 | 82 | 87.56 | 90.18 | 91.74 | 92.42 | 93.72 |
| 18 | Middle-P-T | 58.4 | 61.65 | 56.64 | 59.9 | 62.41 | 61.9 | 61.15 |
| 19 | Plane | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 20 | Car | 88.33 | 85 | 85 | 83.4 | 81.67 | 73.33 | 73.33 |
| 21 | Olive Oil | 77.33 | 56.67 | 53.33 | 48 | 45 | 40 | 40 |
| 22 | Diatom Size Reduction | 95.1 | 95.75 | 96.41 | 94.44 | 93.45 | 91.83 | 81.37 |
| 23 | Gun-Point | 92.67 | 92.67 | 98.67 | 96.33 | 92.67 | 88.67 | 85.33 |
| Average Accuracy% | | 72/81 | 78.39 | 80.33 | 80.90 | 81.02 | 80.22 | 78.94 |

In Table 3, for example for statistical control dataset when $\in = 0.05$ among the 300 time series of experimental dataset the Class of 209 of them is correctly identified which is equal to 69.67%. This process is performed for all data sets and for different values of $\in$. As previously noted, different results of this technique by using different value of similarity threshold in the LCSS indicates the effect of the value of similarity threshold on the result. For example in the case of statistical control dataset, by increasing the value of $\in$ from 0.05 to 0.35, the accuracy of correct recognition of time series class increases from 69.67% to 94.67%. This trend for Gun-Point dataset is initially increcing and then descending, so that its maximum value occurs at $\in = 0.15$. These results show that the value of similarity threshold has very effective on the result of the Query by content technique, and the inappropriate selection of similarity threshold can have adverse effects.

In a general view of the results in Table 3, the accuracy of correct recognition of time series class by applying LCSS method with increase value of similarity threshold from 0.05 to 0.35 in SC, CBF, ECG, Face4, Sweedian, 50words, Distal and Italy's power demand datasets is ascending (i.e., 8 datasets of 23 datasets), in Adiac, Car and Olive Oil datasets is descending (3 datasets of 23 datasets), for GP, Medical, OSU, Beef, Ligthing, Fish, Trace, Ligthing7, Middle-PT, Diatom size reduction and Gun-Point is initially ascending and then is

descending (i.e., 11 datasets from 23 datasets), and eventually this trend for Plane Dataset is initially descending and then is ascending. According to this description and based on the results, the highest accuracy of correct recognition of the time series class for all datasets has occurred in $\in = 0.25$ and is equal to 81.02%.

The interest point of the best-value of similarity threshold (i.e., $\in = 0.25$) is the low accuracy of correct determining time series class in Adiac and Olive Oil datasets which is equal to 42.7% and 45% respectively, which that both have low accuracy and their have worst results among different results of $\in$, So this would be an weakness to the LCSS method.

***Table 4.*** *Accuracy of recognizing the correct class of time series of experimental dataset by the Query by content with DLCSS and* $\in_1 = 0.05$.

| Row | database name | $\in_2$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | **0.20** | **0.25** | **0.30** | **0.40** | **0.50** | **0.6** |
| 1 | Statistical Control | 87.33 | 89 | 90 | 90.67 | 92.67 | 94 |
| 2 | GP | 98 | 97.33 | 97.33 | 97.33 | 96.67 | 96.67 |
| 3 | CBF | 99 | 99.33 | 99.56 | 99.78 | 99.78 | 99.78 |
| 4 | ECG | 87 | 88 | 88 | 85 | 88 | 86 |
| 5 | Face4 | 90.91 | 90.91 | 90.91 | 92.32 | 93.42 | 95.45 |
| 6 | Medical | 58.95 | 58.82 | 60.66 | 60.26 | 61.05 | 62.37 |
| 7 | Sweedian | 79.96 | 79.04 | 80.32 | 82.72 | 84.48 | 85.28 |
| 8 | OSU | 66.94 | 66.94 | 67.95 | 69.01 | 69.01 | 69.23 |
| 9 | Adiac | 81.82 | 82.47 | 83.12 | 83.77 | 83.12 | 82.47 |
| 10 | Beef | 63.33 | 63.33 | 66.67 | 66.67 | 66.67 | 70 |
| 11 | Lighting | 61.64 | 64.38 | 67.12 | 68.49 | 73.97 | 73.98 |
| 12 | Fish | 88.57 | 89.71 | 90.86 | 90.29 | 90.29 | 90.29 |
| 13 | 50words | 70.78 | 71.43 | 73.11 | 74.23 | 76.19 | 77.03 |
| 14 | Trace | 95 | 96 | 96 | 97 | 97 | 97 |
| 15 | Lighhing7 | 61.64 | 64.38 | 67.12 | 68.49 | 72.6 | 73.97 |
| 16 | Distal | 74.75 | 73.5 | 73.5 | 74.25 | 75 | 76 |
| 17 | Italy power demand | 85.52 | 87.85 | 89.99 | 91.64 | 93 | 93.97 |
| 18 | Middle-P-T | 60.65 | 59.15 | 58.9 | 58.65 | 58.65 | 59.4 |
| 19 | Plane | 100 | 100 | 100 | 100 | 100 | 100 |
| 20 | Car | 91.67 | 91.67 | 93.33 | 91.67 | 90 | 88.33 |
| 21 | Olive Oil | 77.33 | 77.33 | 77.33 | 77.33 | 77.33 | 77.33 |
| 22 | Diatom Size Reduction | 95.76 | 95.75 | 95.75 | 95.75 | 96.08 | 96.08 |
| 23 | Gun-Point | 98 | 97.33 | 97.33 | 97.33 | 96.67 | 96.67 |
| Average accuracy % | | 81.0 | 81.4 | 82.4 | 83.0 | 83.8 | 84.3 |

In table 4, the accuracy of the implementation of the Query by content technique by DLCSS method with $\in_1 = 0.05$ and different values for $\in_2$ are presented. As the results show, the accuracy obtained for each dataset is more stable than LCSS's result. For example, the accuracy obtained for Adiac dataset and Olive Oil dataset are more than 80% and %70, respectively. In general overview, the best situation for all datasets is created in a situation where $\in_2 = 0.6$, which is 84.3% and it higher than the best situation obtained by the LCSS method which is 81.02%.

The Query by content technique with DLCSS was implemented again when $\in_1 = 0.10$ and different value of $\in_2$ and the results represent in table 5. As the results show, it is evident that the accuracy obtained for each of the datasets is also. However, contrary to the results in table 4, the accuracy for Adiac and Olive oil dataset are over 68% and about 46%, respectively. In general summary, the best situation for all datasets is created in a state where $\in_2 = 0.6$ which is 83.5% and it is higher than the best accuracy by LCSS method, but in compared with the best situation of the table 4 is less. Therefore, between the different values of similarity threshold for DLCSS method, the best situation is occure at $\in_1 = 0.05$ and $\in_2 = 0.6$.

***Table 5.*** *Accuracy of recognizing the correct class of time series of experimental dataset by the Query by content with DLCSS and* $\in_1 = 0.1$.

| row | database name | $\in_2$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | **0.20** | **0.25** | **0.30** | **0.40** | **0.50** | **0.6** |
| 1 | Statistical Control | 87.33 | 88.33 | 90 | 91.67 | 93.67 | 94 |
| 2 | GP | 98 | 98 | 98 | 98 | 98 | 98 |
| 3 | CBF | 99.33 | 99.67 | 99.78 | 99.78 | 99.78 | 99.89 |
| 4 | ECG | 90 | 89 | 90 | 88 | 87 | 89 |
| 5 | Face4 | 92.05 | 93.18 | 93.18 | 93.18 | 93.18 | 95.45 |
| 6 | Medical | 59.47 | 60.79 | 61.45 | 60.39 | 62.89 | 63.95 |
| 7 | Sweedian | 80.64 | 81.28 | 82.72 | 84.12 | 84.96 | 85.44 |
| 8 | OSU | 69.42 | 68.6 | 68.6 | 69.42 | 70.25 | 69.42 |
| 9 | Adiac | 68.18 | 66.23 | 68.83 | 69.48 | 68.83 | 68.83 |
| 10 | Beef | 73.33 | 76.67 | 76.33 | 73.33 | 73.33 | 73.33 |
| 11 | Lithing | 67.12 | 71.23 | 71.23 | 73.97 | 73.97 | 73.97 |
| 12 | Fish | 90.29 | 89.14 | 90.29 | 90.29 | 89.71 | 88.57 |

| row | database name | $\epsilon_2$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | **0.20** | **0.25** | **0.30** | **0.40** | **0.50** | **0.6** |
| 13 | 50words | 72.83 | 73.67 | 74.23 | 75.35 | 76.47 | 77.31 |
| 14 | Trace | 99 | 99 | 98 | 99 | 98 | 97 |
| 15 | Lithing7 | 67.12 | 71.23 | 71.23 | 73.97 | 73.97 | 73.97 |
| 16 | Distal | 75.25 | 74.5 | 74.25 | 74.75 | 73.5 | 74 |
| 17 | Italy power demand | 87.56 | 89.6 | 90.28 | 92.32 | 92.91 | 94.17 |
| 18 | Middle-P-T | 57.89 | 58.15 | 59.15 | 58.9 | 57.39 | 57.14 |
| 19 | Palne | 100 | 100 | 100 | 100 | 100 | 100 |
| 20 | Car | 90 | 91.67 | 91.67 | 90 | 83.33 | 83.33 |
| 21 | Olive Oil | 46.67 | 46.67 | 46.67 | 46.67 | 46.67 | 46.67 |
| 22 | Diatom Size Reducation | 96.08 | 96.41 | 96.41 | 96.73 | 96.73 | 96.73 |
| 23 | Gun-Point | 98 | 98 | 98 | 98 | 98 | 98 |
| Average Accuracy% | | 81 | 81.5 | 82.2 | 82.8 | 83.1 | 83.5 |

Summary of the best results of implementing the Query by content technique with the DTW, LCSS and DLCSS are presented in Table 6.

It now needs to be checked, is the performance of the DLCSS is better than the DTW? Is the performance of the DLCSS better than the LCSS? For this purpose, pairwise comparison test is used. The zero assumption in this test is the performance of two methods is statistically same, and the one assumption is the performance of two methods is not statistically same. So, If 1% error is tolerable, the interval [1.299, 9.897] is estimated for the accuracy difference between DLCSS and DTW methods and this means that this difference is not zero with 99% confidence and these methods are different in terms of performance, since this difference is positive the performance of the DLCSS is better than the DTW with 99% confidence. Meanwhile, if 10% error is tolerable the interval [0.207, 7.812] is estimated for the accuracy difference between DLCSS ans LCSS and it can be argued that the performance of the DLCSS is better than the LCSS with 90% confidence.

***Table 6.*** *The best Accuracy results of the Query by content technique with the DTW, LCSS and DLCSS methods and their pair comparisons.*

| row | database name | DTW | LCSS | DLCSS | pair comparisons | | | |
|---|---|---|---|---|---|---|---|---|
| | | | **e= 0.25** | **$e_1$=0.05 $e_2$=0.6** | **DLCSS-DTW** | | **DLCSS-LCSS** | |
| 1 | Statistical Control | 97.33 | 93 | 94 | -3.33 | | 1 | |
| 2 | GP | 90.67 | 92.67 | 96.67 | 6 | | 4 | |
| 3 | CBF | 99 | 99.67 | 99.78 | 0.78 | | 0.11 | |
| 4 | ECG | 79 | 90 | 86 | 7 | | -4 | |
| 5 | Face4 | 81.82 | 94.5 | 95.45 | 13.63 | | 0.95 | |
| 6 | Medical | 65.39 | 62.33 | 62.37 | -3.02 | | 0.04 | |
| 7 | Sweedian | 72.16 | 84.62 | 85.28 | 13.12 | | 0.66 | |
| 8 | OSU | 46.28 | 69.83 | 69.23 | 22.95 | | -0.6 | |
| 9 | Adiac | 74.03 | 42.7 | 82.47 | 8.44 | | 39.77 | |
| 10 | Beef | 63.33 | 73 | 70 | 6.67 | | -3 | |
| 11 | Lighting | 68.49 | 75.34 | 73.98 | 5.49 | | -1.36 | |
| 12 | Fish | 75.43 | 81.69 | 90.29 | 14.86 | | 8.6 | |
| 13 | 50words | 66.39 | 73.95 | 77.03 | 10.64 | | 3.08 | |
| 14 | Trace | 100 | 98 | 97 | -3 | | -1 | |
| 15 | Lighting7 | 68.49 | 75.34 | 73.97 | 5.48 | | -1.37 | |
| 16 | Distal | 70.5 | 75.5 | 76 | 5.5 | | 0.5 | |
| 17 | Italy power demand | 93.97 | 91.74 | 93.97 | 0 | | 2.23 | |
| 18 | Middle-P-T | 58.4 | 62.41 | 59.4 | 1 | | -3.01 | |
| 19 | Plane | 100 | 100 | 100 | 0 | | 0 | |
| 20 | Car | 71.67 | 81.67 | 88.33 | 16.66 | | 6.66 | |
| 21 | Olive Oil | 83.33 | 45 | 77.33 | -6 | | 32.33 | |
| 22 | Diatom Size Reduction | 96.41 | 93.45 | 96.08 | -0.33 | | 2.63 | |
| 23 | Gun-Point | 90.67 | 92.67 | 96.67 | 6 | | 4 | |
| Average Accuracy% | | 80.08 | 81.02 | 84.35 | Mean<br>STD | 5.59<br>7.30 | Mean<br>STD | 4.01<br>10.60 |

### *6.2. The K-medoids Clustering Results*

As discussed earlier, the K-medoids clustering technique in this research is used in two steps. In first step, each training dataset is clustered by K-Medoids and the best cluster number and the best cluster representative is selected based on the value of the target unction, then the accuracy index is calculated. In second step, each expremental dataset is grouped based on the

first step results and accuracy of the 2nd step is calculated again. The purpose of this process is to answer these questions:

Question 1: Is the performance of the DLCSS in clustering technique better than the DTW and LCSS performances?

Question 2: Is the performance of the DLCSS in determining the cluster number better than the DTW and LCSS performances?

Question 3: Is the performance of the DLCSS in determining the cluster representative better than the DTW and LCSS performances?

To answer these questions, the clustering technique was implemented on 23 training datasets using the DTW, LCSS and DLCSS in two modes. The first mode is to create 500 initial cluster center and the maximum 200 times displacement of the cluster center. The second mode is to create 500 initial cluster center and the maximum 500 times displacement of the cluster center. The results will be shown in Tables 7 to 12.

First mode**:** Create 500 random initial cluster center and the maximum 200 times displacement of the cluster center

Table 7 shows the results of the implementation of K-medoids clustering technique with the DTW, LCSS and DLCSS methods in the first mode. Based on these results for example for Statistical control dataset, the best result with DTW would be in the cluster number of 6 and with 98.67% accuracy, it means that 98.67% of the time series of this dataset correctly clustered in correct cluster. The best result with LCSS and $\epsilon = 0.25$ is the cluster number of 6 and accuracy of 85.33% and the best result with DLCSS, $\epsilon_1 = 0.05$ and $\epsilon_2 = 0.6$, is 6 for the cluster number and 90.33% accuracy.

Based on these results, the clustering accuracy for all training datasets with DTW is 55.89%, with LCSS is 58.44% and with DLCSS is 62.02%.

***Table 7.*** *K-medoids Clustering of training data set by the DTW, LCSS and DLCSS and Pair comparisons.*

| row | database name | DTW | | LCSS (e=0.25) | | DLCSS ($e_1$=0.05 $e_2$=0.6) | | Pair comparisons | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | K | Accuracy% | K | Accuracy% | k | Accuracy% | DLCSS-DTW | | DLCSS-LCSS | |
| 1 | Statistical Control | 6 | 98.67 | 6 | 85.33 | 6 | 90.33 | -8.34 | | 5 | |
| 2 | GP | 2 | 56 | 4 | 56 | 2 | 56 | 0 | | 0 | |
| 3 | CBF | 3 | 96.67 | 3 | 93.33 | 3 | 96.67 | 0 | | 3.34 | |
| 4 | ECG | 3 | 63 | 2 | 73 | 2 | 72 | 9 | | -1 | |
| 5 | Face4 | 5 | 75 | 6 | 87.5 | 5 | 87.5 | 12.5 | | 0 | |
| 6 | Medical | 8 | 33.86 | 7 | 37.01 | 8 | 37.79 | 3.93 | | 0.78 | |
| 7 | Sweedian | 14 | 51.8 | 13 | 65.4 | 14 | 65.4 | 13.6 | | 0 | |
| 8 | OSU | 4 | 41.5 | 8 | 51 | 8 | 51.5 | 10 | | 0.5 | |
| 9 | Adiac | 32 | 41.54 | 32 | 36.16 | 39 | 46.15 | 4.61 | | 9.99 | |
| 10 | Beef | 7 | 43.33 | 7 | 46.67 | 4 | 43.33 | 0 | | -3.34 | |
| 11 | Lighting | 6 | 58.57 | 7 | 55.71 | 7 | 60 | 1.43 | | 4.29 | |
| 12 | Fish | 6 | 55.43 | 8 | 73.14 | 7 | 82.29 | 26.86 | | 9.15 | |
| 13 | 50words | 54 | 42.22 | 45 | 47.78 | 48 | 52.67 | 10.45 | | 4.89 | |
| 14 | Trace | 3 | 78 | 2 | 52 | 4 | 65 | -13 | | 13 | |
| 15 | Lighting7 | 6 | 55.71 | 7 | 57.14 | 7 | 58.57 | 2.86 | | 1.43 | |
| 16 | Distal | 3 | 59 | 3 | 68.35 | 3 | 68.35 | 9.35 | | 0 | |
| 17 | Italy power demand | 3 | 73.14 | 3 | 65.67 | 3 | 68.66 | -4.48 | | 2.99 | |
| 18 | Middle-P-T | 2 | 55.85 | 2 | 55.85 | 2 | 55.85 | 0 | | 0 | |
| 19 | Plane | 7 | 100 | 7 | 100 | 7 | 100 | 0 | | 0 | |
| 20 | Car | 6 | 56.67 | 5 | 71.67 | 4 | 78.33 | 21.66 | | 6.66 | |
| 21 | Olive Oil | 4 | 86.67 | 3 | 60 | 4 | 83.33 | -3.34 | | 23.33 | |
| 22 | Diatom Size Reduction | 4 | 100 | 4 | 100 | 4 | 100 | 0 | | 0 | |
| 23 | Gun-Point | 2 | 56 | 4 | 56 | 2 | 56 | 0 | | 0 | |
| Average accuracy% | | | 55.89 | | 58.44 | | 62.02 | Mean | 4.22 | Mean | 3.52 |
| | | | | | | | | STD | 9.09 | STD | 5.87 |

K: Expected cluster number from the clustering process

Note: bold number means the correct cluster number and underline number is cluster number with 1 error.

The paired comparison test on the accuracy results in Table 7 is used to answer the first question. If 10% error is tolerable the interval [0.292, 8.151] is estimated for the performance difference between the DLCSS and DTW. This means that this difference with 90% confidence isn't zero, so it can be claimed that the performance of DLCSS is better than the DTW with 90% confidence. If 1% error is tolerable, the interval [0.71, 6.973] is estimated for the performance difference between the DLCSS and the LCSS, so it can be claimed that the performance of DLCSS is better than the LCSS with 99% confidence.

***Table 8.*** *Summary of the number of correctly detects the cluster number for 23 datasets*

| row | Description | DTW | LCSS | DLCSS |
|---|---|---|---|---|
| 1 | Number of Correct predictions of cluster number | 7 | 7 | 13 |
| 2 | Number of predictions of cluster number with 1 error | 8 | 4 | 4 |

To answer the second question referred to the results presented in Table 8, the DTW determines the correct number of clusters for 7 datasets, and this number for the LCSS and DLCSS are 7 and 13, respectively. In general, the DLCSS has the best performance in this area.

After clustering the training datasets and determining the best cluster number and cluster representatives for each of them, the expremental datasets is grouped. These results are present in Table 9. Based on these results and for example for Statistical Control dataset, the time series of experimental dataset can be grouped by the DTW, LCSS and DLCSS with 95.33% accuracy, 84% accuracy and 86.33% accuracy, respectively. In general, for all dataset and by useing the best cluster number and cluster representatives obtained from the first step, the accuracy of grouping by the DTW, LCSS and DLCSS of experimental dataset is 62.35%, 62.64% and 64.91% respectively.

***Table 9.*** *Grouping the Experimental data with cluster centers obtainng from training data clustering and pair comparison.*

| row | database name | DTW | LCSS | DLCSS | pair comparison | | | |
|---|---|---|---|---|---|---|---|---|
| | | Accuracy% | Accuracy% | Accuracy% | DlCSS-DTW | | DLCSS-LCSS | |
| 1 | Statistical control | 95.33 | 84 | 86.33 | -9 | | 2.33 | |
| 2 | GP | 48 | 46.67 | 48 | 0 | | 1.33 | |
| 3 | CBF | 93.33 | 91.33 | 91.33 | -2 | | 0 | |
| 4 | ECG | 60 | 71 | 65 | 5 | | -6 | |
| 5 | Face4 | 52.27 | 85.23 | 87.5 | 35.23 | | 2.27 | |
| 6 | Medical | 30 | 28.29 | 33.03 | 3.03 | | 4.74 | |
| 7 | Sweedian | 52.8 | 67.2 | 67.84 | 15.04 | | 0.64 | |
| 8 | OSU | 35.91 | 44.22 | 44.63 | 8.72 | | 0.41 | |
| 9 | Adiac | 34.02 | 31.46 | 38.62 | 4.6 | | 7.16 | |
| 10 | Beef | 46.67 | 50 | 46.67 | 0 | | -3.33 | |
| 11 | Lighting | 53.43 | 49.32 | 60.27 | 6.84 | | 10.95 | |
| 12 | Fish | 58.86 | 70.29 | 82.86 | 24 | | 12.57 | |
| 13 | 50words | 41.76 | 46.38 | 46.72 | 4.96 | | 0.34 | |
| 14 | Trace | 72 | 48 | 63 | -9 | | 15 | |
| 15 | Lighting7 | 53.43 | 52.06 | 54.8 | 1.37 | | 2.74 | |
| 16 | Distal | 73.25 | 77.25 | 77.25 | 4 | | 0 | |
| 17 | Italy power demand | 73.86 | 63.46 | 65.69 | -8.17 | | 2.23 | |
| 18 | Middle-P-T | 61.16 | 61.16 | 61.16 | 0 | | 0 | |
| 19 | Plane | 99.05 | 99.05 | 100 | 0.95 | | 0.95 | |
| 20 | Car | 51.67 | 48.33 | 63.33 | 11.66 | | 15 | |
| 21 | Olive Oil | 86.67 | 56.67 | 80.33 | -6.34 | | 23.66 | |
| 22 | Diatom Size Reduction | 84.64 | 96.73 | 94.77 | 10.13 | | -1.96 | |
| 23 | Gun-Point | 48 | 46.67 | 48 | 0 | | 1.33 | |
| Average Accuracy% | | 62.35 | 62.64 | 64.91 | Mean<br>STD | 4.39<br>10.30 | Mean<br>STD | 4.02<br>6.97 |

To answer the third question, the paired comparison based on the result in Table 9 is used. If 10% error is tolerable the interval [0.698, 8.087] is estimated for the performance difference between the DLCSS and DTW. This means that this difference is not zero with 90% confidence, so these methods are different in terms of performance and it can be argued that the performance of DLCSS is better than the performance of DTW with 90% confidence. Meanwhile if 2% error is tolerable, the interval [0.381,7.651] is estimated for the difference between the DLCSS and LCSS and it can be argued that the performance of DLCSS is better than the performance of LCSS with 98% confidence.

Second mode: Create 500 random cluster center and allow up to 500 times the center of the cluster to move

Table 10 shows the results of the implementation of K-medoids clustering technique with the DTW, LCSS and DLCSS methods in the first mode. Based on these results and for example for Statistical control dataset, the best result with the DTW would be in cluster number of 6 and with 97.67% accuracy, it means that 97.67% of the time series of this dataset correctly clustered in correct place. The best result with the LCSS and $\in$= 0.25 is cluster number of 6 and 87.33% accuracy and the best result with DLCSS , $\in_1$=0.05 and $\in_2$=0.6 is cluster number of 6 and 91.33% acuuracy. Based on these results, clustering accuracy for all training datasets with the DTW is 56.24%, with LCSS is 58.19% and with DLCSS is 60.61%.

***Table 10.*** *K-medoids Clustering of training data set with the DTW, LCSS and DLCSS and Pair comparisons.*

| row | database name | DTW | | LCSS (e=0.25) | | DLCSS ($e_1$=0.05 $e_2$=0.6) | | Pair comparisons | |
|---|---|---|---|---|---|---|---|---|---|
| | | K | Accuracy% | K | Accuracy% | k | Accuracy% | DLCSS-DTW | DLCSS-LCSS |
| 1 | Statistical Control | 6 | 97.67 | 6 | 87.33 | 6 | 91.33 | -6.34 | 4 |
| 2 | GP | 2 | 56 | 4 | 56 | 2 | 56 | 0 | 0 |
| 3 | CBF | 3 | 96.67 | 3 | 96.67 | 3 | 96.67 | 0 | 0 |
| 4 | ECG | 2 | 60 | 2 | 73 | 2 | 72 | 12 | -1 |
| 5 | Face4 | 3 | 66.67 | 5 | 83.33 | 5 | 87.5 | 20.83 | 4.17 |
| 6 | Medical | 6 | 32.29 | 6 | 31.76 | 8 | 32.28 | -0.01 | 0.52 |
| 7 | Sweedian | 19 | 52.4 | 13 | 66.4 | 13 | 65 | 12.6 | -1.4 |

| row | database name | DTW | | LCSS (e=0.25) | | DLCSS ($e_1$=0.05 $e_2$=0.6) | | Pair comparisons | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | K | Accuracy% | K | Accuracy% | k | Accuracy% | DLCSS-DTW | | DLCSS-LCSS | |
| 8 | OSU | 5 | 43 | 10 | 47.5 | 7 | 48 | 5 | | 0.5 | |
| 9 | Adiac | 30 | 42.31 | 33 | 37.17 | 34 | 46.15 | 3.84 | | 8.98 | |
| 10 | Beef | 6 | 43.33 | 7 | 46.67 | 4 | 43.33 | 0 | | -3.34 | |
| 11 | Lithing | 6 | 55.71 | 10 | 60 | 8 | 57.1 | 1.39 | | -2.9 | |
| 12 | Fish | 9 | 59.57 | 6 | 72.57 | 7 | 82.86 | 23.29 | | 10.29 | |
| 13 | 50words | 48 | 44.89 | 45 | 48.22 | 45 | 50 | 5.11 | | 1.78 | |
| 14 | Trace | 3 | 78 | 2 | 52 | 4 | 71 | -7 | | 19 | |
| 15 | Lithing7 | 6 | 55.71 | 7 | 55.71 | 8 | 55.71 | 0 | | 0 | |
| 16 | Distal | 3 | 59 | 3 | 67.63 | 3 | 61.15 | 2.15 | | -6.48 | |
| 17 | Italy power demand | 3 | 73.14 | 3 | 67.16 | 3 | 68.66 | -4.48 | | 1.5 | |
| 18 | Middle-P-T | 2 | 55.85 | 2 | 55.85 | 2 | 55.85 | 0 | | 0 | |
| 19 | Palne | 7 | 100 | 7 | 100 | 7 | 100 | 0 | | 0 | |
| 20 | Car | 5 | 56.67 | 4 | 70 | 4 | 73.33 | 16.66 | | 3.33 | |
| 21 | Olive Oil | 4 | 86.67 | 3 | 65 | 4 | 83.33 | -3.34 | | 18.33 | |
| 22 | Diatom Size Reducation | 4 | 100 | 4 | 100 | 4 | 100 | 0 | | 0 | |
| 23 | Gun-Point | 2 | 56 | 4 | 56 | 2 | 56 | 0 | | 0 | |
| Average accuracy% | | | 56.24 | | 58.19 | | 60.61 | Mean<br>STD | 3.55<br>8.14 | Mean<br>STD | 2.49<br>6.24 |

K: Expected cluster number from the clustering process

Note: bold number means the correct cluster number and underline number is cluster number with 1 error.

To answer first question, the paired comparison test based on the results in Table 10 is used. If 5% error is tolerable the interval [0.03, 7.074] is estimated for the performance difference between the DLCSS and DTW method. This means that this difference with 95% confidence isn't zero, so it can be claimed that the performance of the DLCSS is better than the DTW with 95% confidence. If 10% error is tolerable the interval [0.253, 4.728] is estimated for the performance difference between the DLCSS and LCSS, and also it can be claimed that the performance of the DLCSS is better than the LCSS with 90% confidence.

***Table 11.*** *Summary of the situation correctly detects the number of clusters for 23 datasets*

| row | Description | DTW | LCSS | DLCSS |
|---|---|---|---|---|
| 1 | Number of Correct predictions of cluster number | 8 | 7 | 10 |
| 2 | Number of predictions of cluster number with 1 error | 7 | 4 | 7 |

To answer the second question referred to the results presented in Table 11, the DTW determines the correct number of clusters for 8 datasets, and this number for the LCSS and DLCSS are 7 and 11, respectively. In general, the DLCSS has the best performance in this area.

After cluster training datasets and determining the best cluster number and cluster representatives for each of them, the expremental datasets is grouped. These results are presented in Table 12. Based on these results and for example for Statistical Control dataset, time series of the experimental dataset can be grouped by the DTW, LCSS and DLCSS with 97.67% accuracy, 87.33% accuracy and 91.33% accuracy, respectively. In general, for all dataset and by useing the best cluster number and cluster representatives obtained from the first step, the accuracy of grouping by the DTW, LCSS and DLCSS of experimental dataset is 62.55%, 62.22% and 64.24% respectively.

To answer the third question the paired comparison test based on the results in table 12 is used. If 10% error is tolerable the interval [0.386, 9.529] is estimated for the performance difference between the DLCSS and DTW. This means that this difference is not zero with 90% confidence, so these methods are different in terms of performance and it can be argued that the performance of DLCSS is better than the performance of DTW with 90% confidence. Meanwhilee if 2% error is tolerable, the interval [0.06,7.726] is estimated for the difference between the DLCSS and LCSS and it can be argued that the performance of DLCSS is better than the performance of LCSS with 98% confidence.

***Table 12.*** *Grouping the Experimental data sets with cluster centers obtainng from training data clustering and pair comparison.*

| row | database name | DTW | LCSS | DLCSS | pair comparison | |
|---|---|---|---|---|---|---|
| | | Accuracy% | Accuracy% | Accuracy% | DlCSS-DTW | DLCSS-LCSS |
| 1 | Statistical control | 96 | 85.57 | 87.33 | -8.67 | 1.76 |
| 2 | GP | 48 | 46.67 | 48 | 0 | 1.33 |
| 3 | CBF | 93.33 | 91 | 91.33 | -2 | 0.33 |
| 4 | ECG | 54 | 71 | 65 | 11 | -6 |
| 5 | Face4 | 44.32 | 86.34 | 87.5 | 43.18 | 1.16 |
| 6 | Medical | 32.76 | 26.45 | 28.55 | -4.21 | 2.1 |
| 7 | Sweedian | 55.04 | 63.84 | 64.64 | 9.6 | 0.8 |
| 8 | OSU | 35.54 | 39.67 | 41.74 | 6.2 | 2.07 |
| 9 | Adiac | 37.85 | 35.29 | 43.22 | 5.37 | 7.93 |
| 10 | Beef | 46.67 | 50 | 46.67 | 0 | -3.33 |
| 11 | Lighting | 51.43 | 50.49 | 52.06 | 0.63 | 1.57 |

| row | database name | DTW | LCSS | DLCSS | pair comparison | | | |
|---|---|---|---|---|---|---|---|---|
| | | Accuracy% | Accuracy% | Accuracy% | DlCSS-DTW | | DLCSS-LCSS | |
| 12 | Fish | 52 | 72.57 | 80.57 | 28.57 | | 8 | |
| 13 | 50words | 40.66 | 45.49 | 49.23 | 8.57 | | 3.74 | |
| 14 | Trace | 72 | 48 | 63 | -9 | | 15 | |
| 15 | Lighting7 | 52.01 | 49.32 | 52.06 | 0.05 | | 2.74 | |
| 16 | Distal | 73.25 | 78.25 | 74.75 | 1.5 | | -3.5 | |
| 17 | Italy power demand | 73.86 | 64.34 | 65.69 | -8.17 | | 1.35 | |
| 18 | Middle-P-T | 61.16 | 61.16 | 61.16 | 0 | | 0 | |
| 19 | Plane | 99.05 | 99.05 | 100 | 0.95 | | 0.95 | |
| 20 | Car | 43.33 | 51.67 | 70 | 26.67 | | 18.33 | |
| 21 | Olive Oil | 86.67 | 56.67 | 80.33 | -6.34 | | 23.66 | |
| 22 | Diatom Size Reduction | 84.64 | 91.83 | 94.77 | 10.13 | | 2.94 | |
| 23 | Gun-Point | 48 | 46.67 | 48 | 0 | | 1.33 | |
| Average Accuracy% | | 62.55 | 62.22 | 64.24 | Mean<br>STD | 4.96<br>12.74 | Mean<br>STD | 3.96<br>6.91 |

# 7. Conclusion

In this research, *a new method for measuring the similarity of time series based on logic and characteristics of the LCSS method is presented which uses two similarity thresholds that named Developed Longest Common Subsequence (DLCSS).* The reasons for using two similarity thresholds in the proposed method are firstly, high felactuation in the implementation of the query by content technique by the LCSS method, secondly, the low accuracy in determining the cluster number of datasets and the accuracy in assigning time series to the right clusters, thirdly, the existence of concepts such as Compactness and Separation in the basic concepts of clustering. In DLCSS method, smaller similarity threshold is the basis for the recognition of the definite similarity between two data and larger similarity threshold as the basis for the recognition of the conditional similarity of the two data. According to the investigations, the best value for them are $\epsilon_1 = 0.05$ and $\epsilon_2 = 0.60$, respectively. By implementation the Query by content technique with the DLCSS, LCSS and DTW method, it was determined that the accuracy of the correct determination of the time series class for the 23 data sets was 84.35%, 81.02% and 80.08% respectively, which that DLCSS mthod has higher accuracy and good stability in results and low error.

In the K-Medoids clustering technique, the accuracy of the clustering of the training datasets with the creation of 500 randomly selected cluster centers and the possibility of 200 displacement of the cluster center with the DTW, LCSS and DLCSS were 55.89%, 58.44% and 62.02% respectively. A pairwise comparison test showed that, it can be claimed that the performance of DLCSS is better than the DTW and LCSS with 95% confidence and 99% confidence respectively. By using the cluster number and cluster representation obtained from the first step, the experimental dataset was grouped with the DTW, LCSS and DLCSS, which have the accuracy of 62.35%, 62.64% and 64.91% respectively. By using pairwise comparison tests, it can be claimed that the DLCSS has better performance in determining the clusters number and cluster representatives than DTW and LCSS with 90% and 95% confidence, respectively. Meanwhile, this clustering process was performed once again by creating 500 randomly selected cluster centers and the possibility of 500 cluster displacements, which shows that DLCSS is superior to DTW and LCSS.

In general, it can be claimed that the DLCSS has a better performance in time series data mining compared to the performance of DTW and LCSS with at least 90% confidence.

# References

[1] Morris, B. & Trivedi, M. (2009), Learning trajectory patterns by clustering: experimental studies and comparative evaluation, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09), pp. 312–319.

[2] Fu, T. C. (2011). A review on time series data mining. Engineering Applications of Artificial Intelligence, 24 (1), pp 164-181.

[3] Keogh, E. & Kasetty, S. (2003). on the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and Knowledge Discovery, 7 (4), pp 349–371.

[4] Sangeeta, R. & Geeta, S. (2012). Recent Techniques of Clustering of Time Series Data: A Survey. International Journal of Computer Applications, 52 (15), pp 1-9.

[5] Lin, J. Vlachos, M. Keogh, E. & Gunopulos, D. (2004). Iterative Incremental Clustering of Time Series. International.

[6] Liao, T. W. (2005). Clustering of time series data: a survey. Pattern Recognition, 38 (11), pp 1857-1874. Conference on Extending Database Technology, Advances in Database Technology- EDBT 2004, pp. 106-122.

[7] Lin, J. Keogh, E. Lonardi, S. & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. DMKD '03 Proceedings of the 8th ACM SIGMOD Workshop on Research issues in data mining and knowledge discovery, pp 2-11.

[8] Aghabozorgi, S. Seyed Shirkhorshidi, A. & Wah, T. Y. (2015). Time-series clustering- A decade review. Information Systems, 53, pp 16-38.

[9] Aghabozorgi, S. Wah, T. Y. Herawan, T. Jalab, H. Shaygan, M. A. & Jalali, A. R. (2014). A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique. The Scientific World Journal, 2014, p562194.

[10] Chen, L. & Ng, R. (2004). On the marriage of Lp-norms and edit distance. VLDB '04 Proceedings of the Thirtieth international conference on very large data bases, 30, pp 792-803.

[11] Esling, P. & Agon C. (2012). Time-Series Data Mining. ACM Computing Surveys, 45 (1), pp. 1-34.

[12] Yi, B. K. & Faloutsos, C. (2000). Fast Time Sequence Indexing for Arbitrary Lp Norms. VLDB '00 Proceedings of the 26th International Conference on Very Large Data Bases, pp 385-394.

[13] Moller-Levet, C. S. Klawonn, F. Cho, K-H. & Wolkenhauer, O. (2003). Fuzzy Clustering of Short Time-Series and Unevenly Distributed Sampling Points. International Symposium on Intelligent Data Analysis, Advances in Intelligent Data Analysis V, pp 330-340.

[14] Berndt, D. J. & Clifford, J. (1994). Using Dynamic Time Warping to find patterns in time series. AAAIWS'94 Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, pp 359-370.

[15] Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR, 163 (4), pp 845–848.

[16] Vlachos, M. Gunopulos, D. & Kollios, G. (2002). Discovering similar multidimensional trajectories. Proceedings 18th International Conference on Data Engineering, pp 673-684.

[17] Chen, L. Ozsu, M. T. & Oria, V. (2005). Robust and fast similarity search for moving object trajectories. SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp 491-502

[18] Vlachos, M. & Gunopulos, D. (2004). Indexing time series under condition of noise. Data mining in time series database: Series in machine perception and artificial intelligence- World Scientific Publishing, 57, pp 67-100.

[19] Vasimalla, K. (2014). A Survey on Tim Series Data Mining. International Journal of Innovative Research in Computer and Communication Engineering, 2 (5), pp 170-179.

[20] Gorbenko, A. & Popov, V. (2012). The Longest Common Subsequence Problem. Advanced Studies in Biology, 4 (8), pp 373-380.

[21] Zhang, Z. Huang, K. & Tan, T. (2006). Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance Scenes. 18th International Conference on Pattern Recognition, 3, pp 1135-1138.

[22] Grabusts, P. & Borisov, A. (2009). Clustering Methodology for Time Sesies Mining. Scientific Journal of RIGA Technical University, computer science, Information technology and management science, 40 (1), pp 81-86.

[23] Ozkan, I. & Turksen, B. (2015). Fuzzy Longest Common Subsequence Matching with FCM. ArXiv.

[24] Gorecki, T. (2014). Using derivatives in a longest common subsequence dissimilarity measure for time series classification. Pattern Recognition Letters, 45 (1), pp. 99–105.

[25] Aghabozorgi, S. & Wah, T. Y. (2014). Effective Clustering of Time-Series Data Using FCM. International Journal of Machine Learning and Computing, 4 (2), pp 170-176.

[26] Lines, J. & Bagnall, A. (2015). Time series classification with ensembles of elastic distance measures. Data Mining Knowledge Discovery, 29 (3), pp 565–592.

[27] Tsai, Y. T. (2003). The constrained longest common subsequence problem. Information Processing Letters, 88 (4), pp 173–176.

[28] Sankoff, D. (1972). Matching Sequences Under Deletion. Insertion Constraints. Proceeding National Academy of Sciences, 69 (1), pp 4-6.

[29] Smith, T. F. & Waterman, M. S. (1981). Identification of Common Molecular Subsequences. Journal of Molecular Biology, 147 (1), pp 195-197.

[30] Amihood, A. Gotthilf, Z. & Shalom, B. R. (2010). Weighted LCS. Journal of Discrete Algorithms, 8 (3), pp 273–281.

[31] Guoa, Y.-P. Pengb, Y.-H. & Yanga, C.-B. (2013). Efficient Algorithms for the Flexible Longest Common Subsequence Problem with sequential sub-string constraints. Journal of Complexity, 29, pp. 44–52.

[32] Cheng, k-Y. Huang, K-S. Yanga, C.-B. & Ann, H-Y. (2013). The Longest Common Subsequence Problem with the Gapped Constriant. The 30th Workshop on Combinatorial Mathematics and Computation Theory, pp 37-42.

# Biography

**Gholamreza Soleimany**: Bachelor of Industrial Engineering from Isfahan University of Technology, Isfahan, IRAN, 1999. Master of Industrial Engineering from Sharif University of Technology, Tehran, IRAN, 2002. Ph.d. student of Industiral Engineering of Yazd University, Yazd, IRAN, since 2015

**Masoud Abessi**: Ph.d. Assistant Professor of Industrial Engineering, Yazd University, Yazd, IRAN. He has published in many journals including IEEE Engineering Management, European Journal of Operation Research and published a book. He has been visiting professor at University of Scranton, U.S.A, University Malaya, Kulala Lumpur, and Tehran University. He has graduated from Clemson University, SC, 1991.