Highlighting Prominent Features for Size Reduction in Time Series Data using Clustering Techniques

1st Anupama Jawale
Research Scholar, PG Department of Computer Science,
SNDT University
Mumbai – 400049 , India
anupama.jawale26@gmail.com

2nd Ganesh Magar
Head, PG Department of Computer Science,
SNDT University
Mumbai – 400049 , India
drgmmagar@gmail.com

***Abstract*—There exist many techniques for feature selection and reduction to reduce dimensions of the large sensor dataset. For real time data processing, compressed and prominent feature of highest significance is desirable for efficient way of resource optimization and computation cost reduction. The goal of this research study is to highlight most significant feature of the dataset and to generate compressed time series by highlighting it. The highlighted feature of accelerometer sensor dataset is extracted, and a more compressed form of time series is generated using statistical and clustering methods like k-means, Partition around Medoids (PAM), Max-Value, 95% Confidence Interval values and Ceil Function calculations. As a result, around 80 % reduction in dataset with the similar pattern as of original time series is achieved. The original time series is compared with generated output series using Dynamic Time Warping method, where, we have obtained normalized error distance of 0.02. (Accuracy 98%)**

Keywords—Accelerometer, Size reduction, Max, K-Means, Dynamic Time Warping, Partition around Medoids.

I. INTRODUCTION

Sensor based devices produce vast amount of data in short time duration. However, processing capacities limited for this type of devices. Sensor based devices are restricted in terms of power, memory and speed. Widely used sensor data streams are vibration sensors, temperature sensors, pressure and level sensors etc. Sensors' embedded systems do not have high speed processors, or graphics processing units, they work on low and limited resources, low cost wireless communications such as Bluetooth Low Energy, wireless networks etc [1]. Real time generation of this data has given rise to challenges pertaining to computational cost, memory and resources requirement and speed. There exist many techniques for feature selection and reduction to reduce dimensions of the sensor dataset. Dimensionality reduction of dataset refers to feature extraction and selection techniques, whereas compression techniques deal with size reduction technique. Size reduction of time series data has its own set of algorithms. However, for the real time series of data points, the size reduction algorithm with faster computation speed is required. One common approach for size reduction of set of data points is clustering algorithm. Popular clustering algorithms like Hierarchical clustering, k means, Partition around Medoids (PAM), Density-based spatial clustering of applications with noise (DBSCAN), Clustering Large Applications (CLARA) are used for size reduction in various areas such as Time series database storage [2], Approximation of time series [3] and so on. Size reduction and clustering also act as a basis for quantization and approximation of time series data. Multiple level clustering [4], Symbolic approximation [5] are certain techniques adopted by researchers for size reduction. Clustering techniques reduce the dataset size by representing a cluster of similar points with single center. With this phenomenon, we have implemented various statistical and clustering, methods on time series dataset to reduce its size as well as highlight its prominent features that are further used for comparison and pattern analysis.

For comparison of time series, techniques like cross correlation, Euclidean Distance between two series, Dynamic time warping are used. In pattern analysis and anomaly detection, of real time time-series data, similarity measurement is done by the method of dynamic time warping [6]. Researchers have used this technique for accelerometer time series processing and comparison [7]–[9], ideal for finding shape symmetry in two waveforms.

In this research study, we have proposed five different methods to reduce the original time series of accelerometer into smaller time series which retains her original characteristics with highlighted prominent feature based on z-axis vibration of accelerometer data. Further we have validated the output of all these

978-1-6654-9291-1/22/$31.00

compressed time series by comparing them with original time series using Dynamic Time Warping.

The main contributions of this research study are

- Novel approach of sliding window processing on a time series ideal for real time processing.
- The size reduction and highlighting of prominent features of dataset achieved by 6 different methods.
- Validation of highlighted reduced time series using dynamic time warping distance calculation and its steps comparison.

This paper is divided further into 4 sections. Section II is about related work and theoretical background, Section III describes proposed methodology whereas Section IV and V describes experimental work and results interpretation, respectively.

II. Related Work

Feature reduction and dimensionality reduction techniques and their mathematical and statistical base can be broadly classified as below [10].

A. Statistical information-based methods – Reducing number of observations based on some statistical information in such a manner that newly calculated set of values (k) represents some set if observations (N).
B. Vector Quantization and Mixture Models – Probability is considered as a classifier to classify each observation. Thus, it works as mapping N number of observations into k number of probability classes. For example, PCA
    - Principal curves, surfaces, and manifolds – Data that does not fit into linear manifold can be reduced by transforming a curve into another dimensional space as a line.
C. Generative Topographic Mapping – The observed data is generated by mapping data points into a dimensional space which is much lower.
    - Self Organizing Maps – SOM are generalized form of vector quantization where each observation is assigned some label corresponding to its closest class.
    - Elastic maps, nets, principal graphs and principal trees – For each data point, there is a mapping between low and high dimensional spaces.

Being an important step of many data mining tasks dimensionality reduction attracts focus of many researchers. This paper tries to address this issue by just keeping prominent features in the time series as described in next section.

D. Dictionary based Methods – This method is based on crumbling a matrix formed by data points. The change of basis is known as dictionary. Below are some popular techniques under this category
    - Nonnegative matrix factorization
    - Tensor analysis and tensor factorization
    - Generalized Singular value decomposition
E. Projection based Methods – This type of methods are primarily focus on projecting original data into subspace with important features. Methods under this category are enlisted below
    - Interesting directions
    - Manifolds

In time series applications, clustering techniques are used for pattern recognition and time series aggregation to construct similar scenario for unknown inputs [11]. Fault detection is another common example to time series data. Vibration sensors are used to capture the faulty equipment. K-means clustering is effectively used to detect faulty vibrations, even from the raw vibration data with noise [12]. Feature extraction followed by Eigen value normalization for clustering data into normal/faulty cluster has achieved at most 94-100% accuracy in prediction. K-means has been successfully implemented to cluster industrial customers for load forecasting. This is used in combination with CatBoost [13]. In another application of k-means on time series, researchers have achieved dataset reduction using Symbolic Aggregate Approximation and Piecewise Aggregate Approximation techniques to improve clustering quality and analysis efficacy [14].

To compare one or more time series in different temporal sequence, cross correlation method fails if the time shift is large. Dynamic Time warping is the method of comparison of such temporal sequences for similarity and pattern or anomaly discovery [15]. New method for similarity measurement from music pieces is proposed in research study [16] that extracts binary chromogram and maps it to tonal centroid representation with quantization. Similarity is measured on the basis of size reduction base similarity metric.

III. METHODOLOGY

The novelty of this research study comes from the concept of substituting highlighted prominent features of a Time Series (TS) data set into original series to generate compressed series. We have carried out this research study in three different stages. First stage is exploratory data analysis where feature selection form the original raw time series data using Principal Component Analysis is performed. Then in second stage, the selected feature is processed using five different methods for size reduction and highlighting it. of Z-Axis Accelerations (ZAcc) is the selected feature from processing, Any variation in vibration value ZAcc indicates strong movement along Z- axis, against gravitational force [7]. In third stage, we have compared compressed series with original time series using dynamic time warping on the basis of normalized distance between them. The methodology is explained in below flowchart (figure 1) and illustrated in depth in later part of this section.

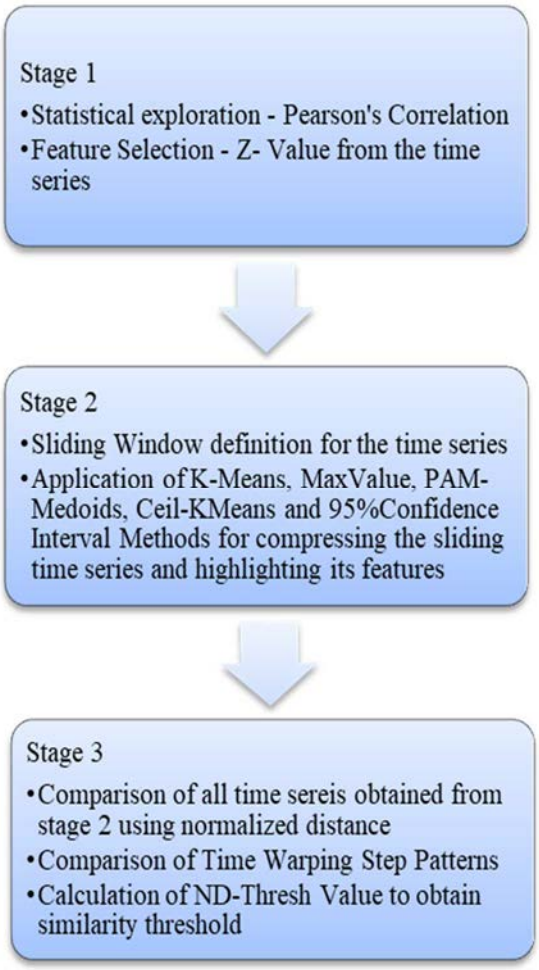

Figure 1: Methodology behind size reduction and comparison techniques

Stage 1 – Exploratory Data Analysis and Feature Selection using Principal Component Analysis. Figure 1 and 2 show graphical representation of Pearson's Correlation Coefficient and Principal Component Analysis respectively, whereas figure 3 explores z-acc value of all time series. Traditional approaches perform feature reduction techniques to reduce the dataset [17]. Pearson's Correlation Coefficient indicates association between two variables. Principal Component Analysis (PCA) is a technique to reduce dimensionality of the data set [18]. It is data exploratory analysis tool. In its standard format, for a dataset on n observations, it defines p number of n dimensional vectors as n X p matrix. From this matrix, we discover a linear combination of maximum variance. This linear combination is given by the equation

$$\sum_{j=1}^{p} a_j x_j = x_a \quad (1)$$

Where a is $a_1,\ a_2, a_3, a_4, \ldots\ldots\ldots\ a_p$

PCA, as its dynamic version can be used if there is a correlation among p variables for vector time series [19]. The accelerometer time series data consists of seven different features, out of these PCA has selected the z-feature as the most relevant feature on the basis on ground truth of vibration movement detection in vertical axis.

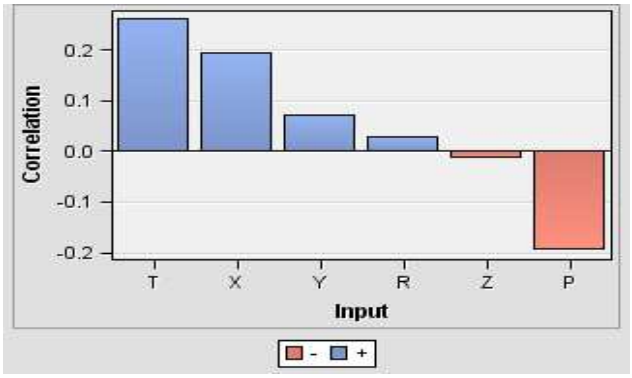

Figure 2 : Pearson's Correlation Coefficient from TS Data

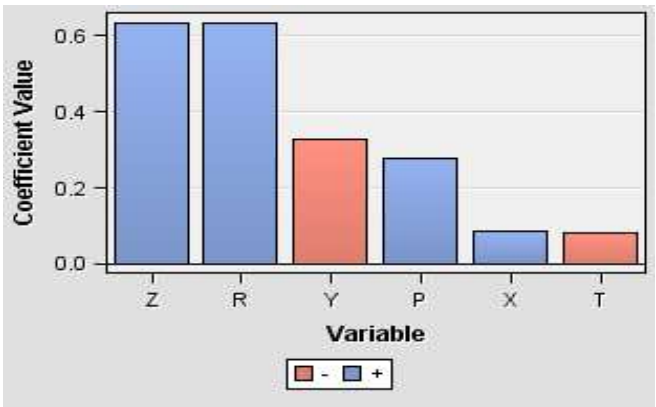

Figure 3 : Principal Component Analysis

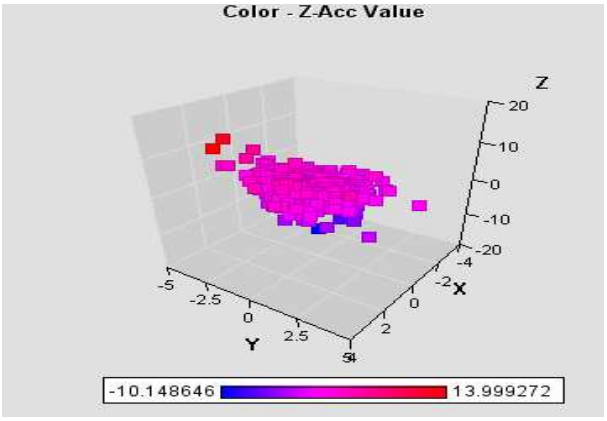

Figure 4 : Exploration of selected feature – ZAcc Value

Stage 2 – Size reduction Methods for highlighting features and reducing size of TS

In this research study we have applied clustering algorithms, statistical generic methods for converting the larger time series into smaller one but representing same features for z-values. The prominent feature value is representative of the original series. The technique is applied on sliding window (Figure 5) of a TS with size formula

$$Window\ size\ T_{\propto} = \frac{n}{1024} \approx 20 \quad (2)$$

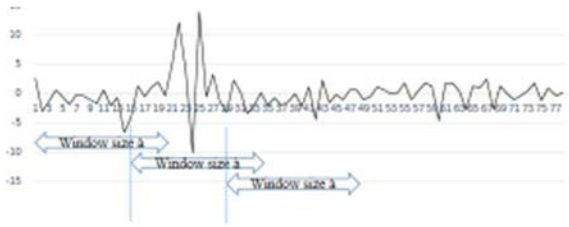

Figure 5 : Sliding window applied to Data point Zacc

The methods of size reduction of TS are on the basis of clustering as well as Max, Confidence Interval statistical functions. These methods are applied to sliding window for faster real time processing. The summary of these methods is shown in the table below.

Table 1 : Summary of Methods used for size reduction of TS

Method	Description
K – Means algorithm (Traditional k-Means) Abbreviation KQ	K-Means algorithm computes clusters by dividing dataset into predefined number of clusters. In this research study to achieve 20 % reduction in data set size, we have taken number of clusters = $\frac{T_{\alpha}}{5}$. K-Means objective function can be given by formula $\sum_{j=1}^{k}\sum_{i=1}^{n}\|x_i^{\,j} - c_j\|^2$ (3) Where, k=Number of clusters, c is centroid of the cluster defined by mean value, x_n is data point at ith case
Ceil K-Means Algorithm (Modified K-means) Abbreviation CKQ	The Accelerometer Z-acc values are very sensitive. The accuracy of the data point can be optimized for grouping purpose. The logic behind Ceil-K-means function is to convert the data point to nearest ceiling value before forming a cluster. Approximation of Ceiling can be defined by factors affecting the error term. $\sum_{j=1}^{k}\sum_{i=1}^{n}\|ceil(x_i)^{j} - c_j\|^2$ (4)
MaxValue Algorithm (Novel Method) Abbreviation MQ	The highest ZAcc value identifies the maximum amount of vibration received against gravity in vertical direction. Identifying top max amplitudes describes the highlighting feature of the window. A time series in given window is considered as a continuous function f(x) where x=ZAcc. The series is finite. If f is continuous on a closed interval [a, b], then there exists minimum and a maximum point. Thus, there are real numbers c and d in [a, b] such that for every x in [a, b], f(x) ≤ f(c) and f(x) ≥ f(d). MaxValue Algorithm is given by the same formula, but for each iteration, respective f(d) value is excluded from the series.
PAM Algorithm (Traditional k-Medoid) Abbreviation PAM	PAM (Partitioning around medoids is another popular algorithm that works similar to K-means. Medoids are the most centrally located points of the cluster. PAM's objective function is given by the formula (3) where cj is medoid of the cluster.
95 % Confidence Interval Algorithm (Novel Method) Abbreviation 95PC	95% Confidence level is a statistical measure. It indicates that if the same set of values is sampled on different intervals and interval estimates are calculated on each occurrence, the resulting intervals would contain the true dataset parameter in approximately 95 % of the cases. A confidence stated at a

	1−α level can be thought of as the inverse of a significance level, α. Here we extracted left, right and mean value from the 95% confidence interval to highlight the original dataset feature. Formula for left, right and mean values is given as follows $95\%Int = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ (4)

Stage 3 – Comparison of highlighted features' TS with original TS

In this stage, dynamic time warping is used for comparison of original and compressed series obtained from stage 2. It is a well-known technique for alignment of temporal series. Temporal series with different time dimensions can either be stretched or compressed to map with other series. For dynamic time warping distance measure has to be defined. In this research study we have used the default Euclidean distance measure. The formula for Euclidean distance is given as follows.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (5)$$

Once the distance measure is formulated, dynamic time warping algorithm can be formulated by finding out the optimized distance between data values [20]. For similarity search, we convert this problem into minimization problem such that,

$$DTW(S,T) = Min_w\left[\sum_{k=1}^{p}\delta(w_k)\right] \quad (6)$$

The next section shows experimental results obtained in this research study.

IV. Experimental Work

The dataset used for the experiments is a Time Series of accelerometer data, on a dashboard of moving Car. This series has 20000 observations with 6 dimensions. Table 2 shows the outcome of size reduction methods in terms of normalized distance, when compared all output series with original time series. Each of the output series contains 80% reduction of the original dataset with 98-100 % accuracy, preserving its highlighted features. The normalized distance tolerance term is around 0.15 is assumed for calculation of NDE (Normalized Distance Error)

Table 2 : Normalized Distance comparison between highlighted feature output series and original series.

WS	KQ NDE	PAM NDE	CKQ NDE	95PC NDE	MQ NDE
10	0.02	0.06	-0.02	0.03	0.04
16	0.16	0.32	0.13	0.16	0.18
20	0.02	0.00	0.02	0.03	0.11
24	0.16	0.32	0.13	0.16	0.18

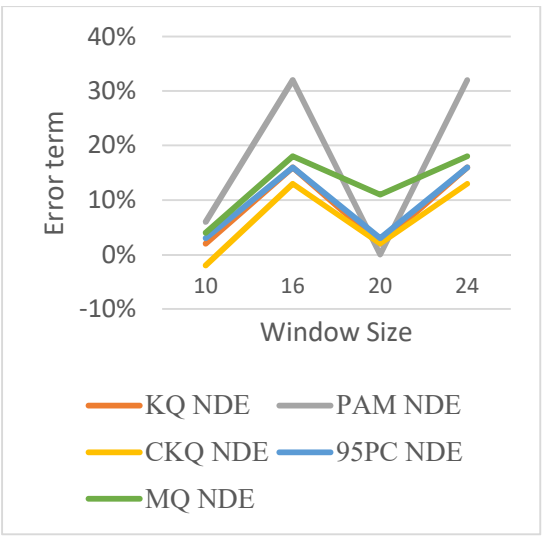

Figure 6: Normalized Distances of highlighted feature series

V. Conclusion and Future Scope

The results illustrated in previous section have shown that feature highlighting using technique K-Means, Ceil K-Means, PAM and 95% Confidence levels is useful for reducing the large size time series to smaller one and proves it as a replica of original series with 98 - 100 % accuracy. Distance calculation between original and reduced series also shows significant similarity in the step patterns of dynamic time warping comparison. This technique would help in continuous real time processing of time series data for pattern detection, anomaly detection etc. In future we would like to work on more robust method of reduction and quantization of time series data detection etc. In future we would like to work on more robust method of reduction and quantization of time series data.

REFERENCES

[1] L. Klus, D. Quezada-Gaibor, J. Torres-Sospedra, E. S. Lohan, C. Granell, and J. Nurmi, "RSS Fingerprinting Dataset Size Reduction Using Feature-Wise Adaptive k-Means Clustering," in 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic, Oct. 2020, pp. 195–200. doi: 10.1109/ICUMT51630.2020.9222458.

[2] T. Pelkonen et al., "Gorilla: a fast, scalable, in-memory time series database," Proc. VLDB Endow., vol. 8, no. 12, pp. 1816–1827, Aug. 2015, doi: 10.14778/2824032.2824078.

[3] F. Cicalese and E. S. Laber, "Information Theoretical Clustering is Hard to Approximate," IEEE Trans. Inform. Theory, pp. 1–1, 2020, doi: 10.1109/TIT.2020.3031629.

[4] S. Aghabozorgi, T. Ying Wah, T. Herawan, H. A. Jalab, M. A. Shaygan, and A. Jalali, "A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique," The Scientific World Journal, vol. 2014, pp. 1–12, 2014, doi: 10.1155/2014/562194.

[5] N. Ruta, N. Sawada, K. McKeough, M. Behrisch, and J. Beyer, "SAX Navigator: Time Series Exploration through Hierarchical Clustering," in 2019 IEEE Visualization Conference (VIS), Vancouver, BC, Canada, Oct. 2019, pp. 236–240. doi: 10.1109/VISUAL.2019.8933618.

[6] Z. Chen and J. Gu, "High-Throughput Dynamic Time Warping Accelerator for Time-Series Classification With Pipelined Mixed-Signal Time-Domain Computing," IEEE J. Solid-State Circuits, pp. 1–1, 2020, doi: 10.1109/JSSC.2020.3021066.

[7] A. Allouch, A. Koubaa, T. Abbes, and A. Ammar, "RoadSense: Smartphone Application to Estimate Road Conditions Using Accelerometer and Gyroscope," IEEE Sensors J., vol. 17, no. 13, pp. 4231–4238, Jul. 2017, doi: 10.1109/JSEN.2017.2702739.

[8] M. R. Carlos, M. E. Aragon, L. C. Gonzalez, H. J. Escalante, and F. Martinez, "Evaluation of Detection Approaches for Road Anomalies Based on Accelerometer Readings—Addressing Who's Who," IEEE Trans. Intell. Transport. Syst., vol. 19, no. 10, pp. 3334–3343, Oct. 2018, doi: 10.1109/TITS.2017.2773084.

[9] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," Pattern Recognition, vol. 44, no. 9, pp. 2231–2240, Sep. 2011, doi: 10.1016/j.patcog.2010.09.022.

[10] C. O. S. Sorzano, J. Vargas, and A. Pascual, "A survey of dimensionality reduction techniques," p. 35.

[11] M. Ma, L. Ye, J. Li, P. Li, R. Song, and H. Zhuang, "Photovoltaic Time Series Aggregation Method Based on K-means and MCMC Algorithm," in 2020 12th IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), Nanjing, China, Sep. 2020, pp. 1–6. doi: 10.1109/APPEEC48164.2020.9220338.

[12] Q. Wang, J. Liu, B. Wei, W. Chen, and S. Xu, "Investigating the construction, training and verification methods of k-means clustering fault recognition model for rotating machinery," IEEE Access, pp. 1–1, 2020, doi: 10.1109/ACCESS.2020.3028146.

[13] C. Zhang, Z. Chen, and J. Zhou, "Research on Short-Term Load Forecasting Using K-means Clustering and CatBoost Integrating Time Series Features," in 2020 39th Chinese Control Conference (CCC), Shenyang, China, Jul. 2020, pp. 6099–6104. doi: 10.23919/CCC50068.2020.9188856.

[14] Y. Shi, T. Yu, Q. Liu, H. Zhu, F. Li, and Y. Wu, "An approach of electrical load profile analysis based on time series data mining," IEEE Access, pp. 1–1, 2020, doi: 10.1109/ACCESS.2020.3019698.

[15] G. Huang, X. Chen, and Y. Chen, "P-P and Dynamic Time Warped P-SV Wave AVA Joint-Inversion With ℓ1-2 Regularization," IEEE Trans. Geosci. Remote Sensing, pp. 1–14, 2020, doi: 10.1109/TGRS.2020.3022051.

[16] T. E. Ahonen, K. Lemstrom, and S. Linkola, "COMPRESSION-BASED SIMILARITY MEASURES IN SYMBOLIC, POLYPHONIC MUSIC," Poster Session, p. 6, 2011.

[17] B. Ghojogh et al., "Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review," p. 15.

[18] V. B.D W. N. ,. Ripley, Modern Applied Statistics with S .

[19] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Phil. Trans. R. Soc. A., vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[20] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," p. 12.