Time Series Segmentation Using Two-Stage Clustering Approach

Kaushik C Shete, Dr. Amar Buchade

Abstract- Time series is a sequence of observations of data points measured over a time interval. Time series segmentation organizes time series into partitions that are having similar characteristics. These segments can be used for analyzing a particular time sequence, dimensionality reduction, removing outliers and for similarity search. Time series reduction is a process of reducing data points from the original time series. Time series reduction will lead to the system to operate on a reduced number of data points. The problem with multiple time series is difficult to analyze. By segmenting this, one can easily analyze. The proposed TS (time series) segmentation method is based on a two-stage clustering approach, where the first stage is responsible for extracting correlated time series and the second stage is responsible for extracting time series based on their reduction or magnitude values. These segments are helpful in stock market analysis, for finding energy consumption patterns, extracting cyclic or trending TS patterns; also it is useful for finding similar time series.

Index Terms- Data Mining, Time Series Analysis, Cluster Analysis, Correlation Analysis, Data Mining, Time Series Segmentation, Dimensionality Reduction.

1 INTRODUCTION

Time series are used in statistics, signal processing, pattern recognition, finance, forecasting. Time series generally have historical data. Time series analysis can be performed in different areas for various purposes like subsequent matching, indexing, clustering, classification, visualization, segmentation, trend analysis, forecasting, etc. Data is measured over a specific time interval like a daily basis or hourly basis for a long period with multiple parameters. Because of having high dimensions and large size users cannot easily visualized time series. A time-series database may have some valuable information. This information can be extracted from pattern discovery. Time series reduction is a process of reducing data points from the original time series. Time series reduction will lead to the system to operate on a reduced number of data points. Time series segmentation is a process of dividing time series into several segments or classes. Segmentation leads to the discovery of patterns or trends. These segments will help to reduce computational power as dimensions reduce. Segmentation can be classified into-

- Vertical Segmentation: In this technique, segments are made on a vertical basis. Segments can be used for extracting patterns like trends, seasonal, cyclic, etc and also can be used to reducing time series. It is usually applied to singular TS data. Figure 1 shows the output of vertical segmentation with six segments as S1 to S6, where S2 & S4 show an upward trend whereas segment S5 shows a downward trending pattern.

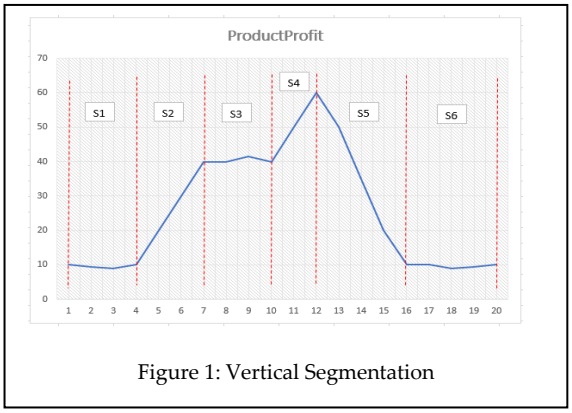

Figure 1: Vertical Segmentation

- Horizontal Segmentation: In this technique, segments are made on a horizontal basis. This technique is used for extracting a similar time series. Using this approach user gets segments, where each segment is set similar time series. Also, users can select some time series from the available graph [5]. Figure 2 shows the output of horizontal segmentation with two segments as S1 & S2. Segmentation is applied to SAS Stock-Market data-set. The dotted line helps the user to determine the separation of segments between time series data.

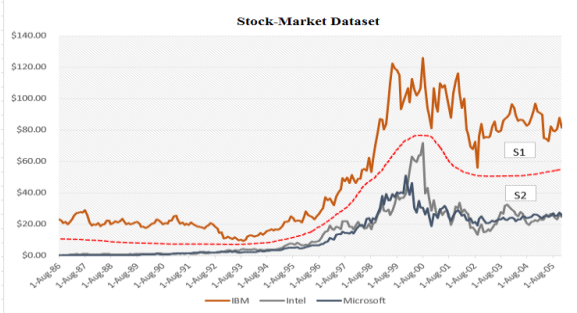

Figure 2: Horizontal Segmentation

- *Kaushik C Shete, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India, shetekaushik@gmail.com*
- *Dr. Amar Buchade, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India, arbuchade@pict.edu*

This paper organized into five sections. Section II includes a literature review. System overview, objectives of system and algorithms for various approaches are covered in section III. Section IV covers the results and analysis part. Section V gives performance measures that are used for comparing the proposed system with existing techniques. In the end, section VI represents the conclusion and future enhancement part.

2 Review of Literature

1. **Segmenting Time Series: A Survey And Novel Approach [1]**

In this, authors have proposed vertical time series segmentation based on Sliding Window And Bottom-up approach. Segmentation of time series is based on a single variable. Also, the authors gave details of existing segmentation techniques like sliding window, top-down, bottom-up. The proposed system is compared with other techniques using various datasets like Radiowaves, ExchangeRates, WaterLevel, SpaceShuttle.

2. **An Algorithmic Method for Segmentation of Time Series: An Overview [2]**

In this, authors have given the information on various segmentation techniques based on piecewise linear representation. The authors have provided algorithms for each technique. Techniques are compared with each other using various datasets like CapacityUtilization, WheatPrice, ECG and StockMarket dataset.

3. **Review on Time Series Data Mining [3]**

It is a review of time series data mining. Authors have given various representation and indexing techniques, Similarity measures, Segmentation, Visualization and Data mining of time series. Also, authors have given information related to time series dimensionality reduction which is helpful to reduce time series data.

4. **Time Series Analysis and Forecasting [4]**

In this, authors have provided information related to time series analysis and forecasting. Authors have provided the classification of patterns like horizontal pattern, trend pattern, seasonal pattern, cyclic pattern, etc. It also provided information about various forecasting techniques for time series data.

5. **SAS Time Series Studio 14.1: Users Guide [5]**

It is software that gives segmentation functionality for multiple time series data. Segmentation is provided by 3 approaches:

- Graphical Query Approach- Using this user can create a segment by selecting the desired time series from the graph. For selecting the desired time series user able to draw a rectangle bar using mouse pointer around the region of interest of time series.
- Hierarchical Query Approach- Using a hierarchical query users can select leaf or level in the hierarchy. Based on the selection of the hierarchy time series will segment.
- Parameter Query Approach- Using this, based on the input level dataset or descriptive statistics parameters user able to divide time series into the segments.

6. **Time Series Clustering – A Decade Review [6]**

It is a review of time series clustering. Authors provided various TS clustering based on- i) Whole-time series clustering ii) Sub-sequence time series clustering iii) Time point clustering. It also provides various clustering algorithms based on Hierarchical, Partitioning, Grid, Density, etc.

7. **Hierarchical Clustering of Time-Series Data Streams [7]**

In this, the authors have proposed the Online Divisive Agglomerative Clustering algorithm for time series clustering. In this system iteratively goes on splitting time series data until stopping criteria met. While splitting time series into 2 parts also it has the ability to aggregate clusters. The proposed system has the ability to split and aggregate the clusters.

8. **Auto-correlation Based Fuzzy Clustering of Time Series [8]**

Pierpaolo D'Urso et.al proposed a fuzzy clustering algorithm for time series clustering based on auto-correlation values for different lags. The proposed system is based on a non-hierarchical clustering approach. A fuzzy C-means clustering procedure gives different advantages like simplicity, speed, and good computational performances. Auto-correlation measures the linear relationship between lagged values of a time series which helps to find out the degree of similarity. Correlation values are helpful to extract information such as trends, cycles and normal patterns from time series data.

9. **Clustering Stationary and Non-stationary Time Series Based on Auto-correlation Distance of Hierarchical and K-means Algorithms [9]**

D. Pratiwi et.al gives a comparison of hierarchical and K-Means clustering using ACF values for clustering time series data with respect to accuracy. The authors have used both simulated and real datasets for the research. Hierarchical and K-means clustering is used because of its characteristics- efficiency, scalability, and simplicity. Authors stated that K-means produce more accuracy than the hierarchical clustering algorithm.

10. **TSrepr - Time Series Representations in R [10]**

TSrepr is a package in R that provides methods for time series representations and several other useful helper

methods and functions. This provides various methods for reducing time series data using R. Methods are based on-

a) Non-data Adaptive: PAA, PIP, DFT
b) Data Adaptive: SAX
c) Model-Based: Mean seasonal profile, Exponential smoothing

These approaches are helpful for reducing time series data into a fixed number of data points.

11. Feature Extraction Methods for Time Series Data in SAS® Enterprise Miner™ [11]

This provides various approaches for time series reduction. It also provides various feature extraction methods for time series data. Feature extraction methods using classical TS analysis are-

a) TS Reduction with Time Intervals
b) Seasonal Analysis
c) Correlation Analysis
d) Seasonal Decomposition

12. Time Series Classification and its Applications [12]

Krisztian Buza gave information about challenges, methods, evaluation protocols and bio-medical applications related to time series classification. The author has covered preprocessing techniques based on FFT (Fast Fourier Transform) and SAX (Symbolic Aggregate Approximation). Gave information about TS classification using similarity-based, feature-based and CNN.

3 Proposed Methodology

The existing segmentation techniques have high complexity and large processing time because of distance-based clustering mechanisms as every TS variable compares with other TS variable. In this paper, time series segmentation is proposed based on a two-stage clustering approach. The first stage is responsible for extracting time series based on correlation analysis whereas the second stage extracts time series based on a single point reduction approach.

In this we have 5 approaches for TS reduction as follows- Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), Singular Value Decomposition (SVD), Line Segment Approximation with Sum (LSAS) and Line Segment Approximation with Mean (LSAM).

3.1 Objective and Scope

To design and implement time series segmentation using a two-stage clustering approach and test its efficiency on different time series data-sets.

Objectives:

- Study different time series reduction and segmentation techniques.
- To implement time series segmentation using correlation analysis and reduction approach.
- To test the efficiency of methodology on different datasets.
- To compare the results of the proposed methodology with other existing segmentation techniques.

Scope:

- Working with a lower hierarchy.

3.2 Architecture

The system is divided into 6 phases- Preprocessing, Correlation analysis, Stage_1 clustering, Reduction phase, Stage_2 clustering, and Segmentation.

Figure 3 shows a system overview of the proposed system. Preprocessing is used for converting raw data into time series data. Correlation analysis is used for extracting lags values. Stage_1 clustering is used for extracting correlated time series. The reduction phase is responsible for converting TS data into a single dimension using available techniques. Stage_2 clustering is applied on reduced values for segmenting time series based on their magnitude. K-means is used for creating clusters. The segmentation phase creates segments on the basis clustering outcome.

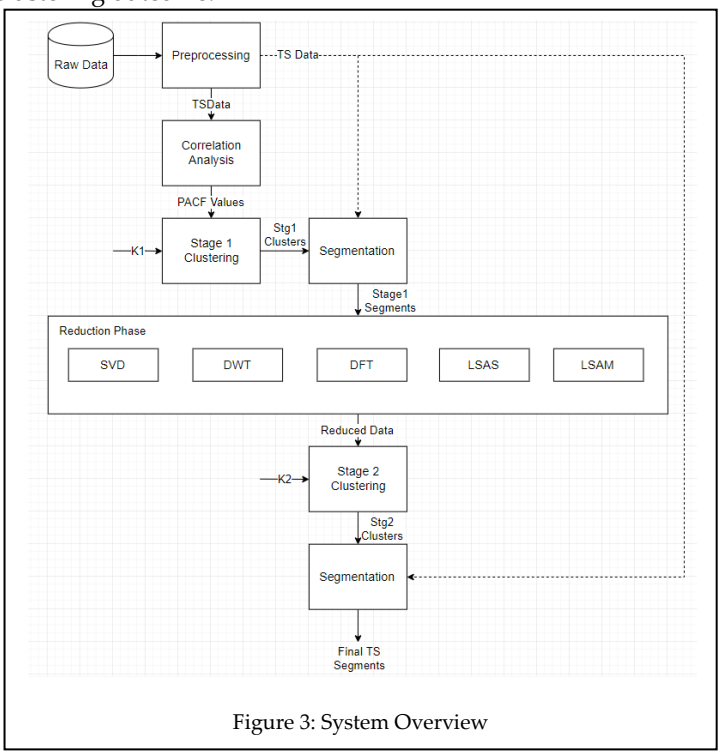

Figure 3: System Overview

3.3 Proposed Algorithm

This section provides a proposed algorithm and mathematical model of the system.

Algorithm 1: Time Series Segmentation

Input : Data, K_1, K_2, $Reduction_{Approach}$

Output : $TSFinal_{Segments}$

1. $TS_{Data} = Preprocess(Data);$
2. **if** $Valid(TS_{Data})$ **then**
3. | $GOTO\ Step6;$
4. **else**
5. | $STOP;$
6. $PACF_{Values} = CorrelationAnalysis(TS_{Data});$
7. $PACF_{Indicators} = FindIndicators(PACF_{Values});$
8. $Stage1_{Clusters} = Clustering(PACF_{Indicators}, K_1);$
9. $Stage1_{Segments} = Segmentation(TS_{Data}, Stage1_{Clusters})$
10. **if** $Reduction_{Approach} = 1$ **then**
11. | $Reduced_{TSData} = SVD_{Reduction}(Stage1_{Segments});$
12. **if** $Reduction_{Approach} = 2$ **then**
13. | $Reduced_{TSData} = DFT_{Reduction}(Stage1_{Segments});$
14. **if** $Reduction_{Approach} = 3$ **then**
15. | $Reduced_{TSData} = LSAS_{Reduction}(Stage1_{Segments});$
16. **if** $Reduction_{Approach} = 4$ **then**
17. | $Reduced_{TSData} = LSAM_{Reduction}(Stage1_{Segments});$
18. **if** $Reduction_{Approach} = 5$ **then**
19. | $Reduced_{TSData} = DWT_{Reduction}(Stage1_{Segments});$
20. $Stage2_{Clusters} = Clustering(Reduced_{TSData}, K_2);$
21. $TSFinal_{Segments} = Segmentation(TS_{Data}, Stage2_{Clusters});$

1. ***Correlation Analysis:*** It will help to extract correlated time series using a partial auto-correlation function.

$$PACF_{1,1} = ACF_1$$

$$PACF_{k,k} = \frac{ACF_k - \sum_{i=1}^{k-1} PACF_{k-1,i} * ACF_{k-i}}{1 - \sum_{i=1}^{k-1} PACF_{k-1,i} * ACF_i}$$

$$PACF_{k,i} = PACF_{k-1,i} - PACF_{k,k} * PACF_{k-1,k-i}$$

For i= 1, 2, 3,…k-1

Where,

$PACF_{k,k}$ = Partial Auto-correlation at lag k w.r.t. time series.

ACF_k = Auto-correlation at lag k w.r.t. time series.

2. ***Reduction Phase:*** Time series reduction is used for reducing each time series variable into a single data point. We have the following 5 approaches for time series reduction-

- **Discrete Fourier Transform**

It converts time series on the basis of the frequency domain.

$$Y = \sum_{i=1}^{N} X(i) e^{-\frac{j\pi n}{N}}$$

Where,

Y = Single point reduced time series.

X = Time series to be reduced.

N = Total number of data points.

- **Singular Value Decomposition**

It is based on matrix decomposition.

$$A = U * D * V$$

Where,

A = Time series matrix of size m*1 that to be reduced.

U = V = Orthogonal matrix of size m*m.

D = Reduced matrix of size 1*1.

- **Discrete Wavelet Transform**

It converts time series based on wavelet function.

$$Y_{(i,j)} = h * Y_{(i-1,j)} + h * Y_{(i-1,j+1)} \cdots$$

Iterate above till i!=0;

Where,

$h = 2^{-1/2}$

i = Resolution level.

j = Point at respective resolution level.

- **Line Segment Approximation with Sum**

Sum of all data points with respective of time series variables.

$$Y = \sum_{i=1}^{N} X(i)$$

Where,

Y = Single point reduced time series.

X = Time series to be reduced.

N = Total number of data points.

- **Line Segment Approximation with Mean**

Average of all data points with respective of time series variables.

$$Y = \frac{\sum_{i=1}^{N} X(i)}{N}$$

Where,

Y = Single point reduced time series.

X = Time series to be reduced.

N = Total number of data points.

3. ***Clustering:*** K_means clustering algorithm is used for stage_1 and stage_2 clustering.
4. ***Segmentation:*** Segmentation is used for making $Stage1_{Segment}$ and $TSFinal_{Segments}$ using stage_1 and stage_2 clusters respectively with input TS data.

4 RESULTS

This section provides a snapshot of the results based on Partial Auto-correlation followed by the Reduction Approach.

- **Using Partial Auto-Correlation and Reduction Approach**

These results we got from stage_1 and stage_2 clustering. Figure 4 is a snapshot of the graphical representation input time series data-set collected from the UCI repository. Figures 5 and 6 show the results of stage_1 and stage_2 clustering respectively. Figure 6 shows segments from stage_1 cluster 2.

Figure 5 clusters shows cyclic, trending, and stationary patterns cluster is equals to 1, 2, and 3 respectively. Figure 6 shows downward and upward trending patterns are classified as final TS Segments after stage_2 clustering.

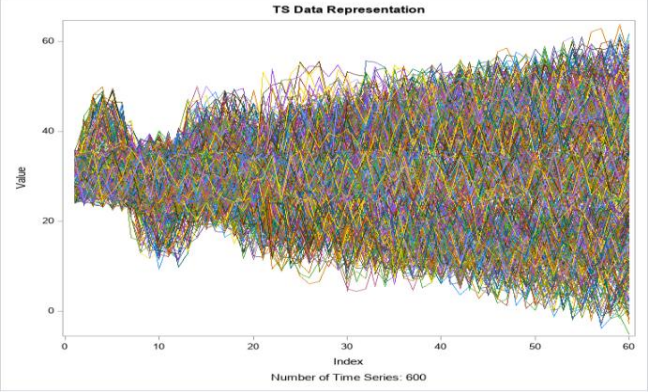

Figure 4: Synthetic Data Representation

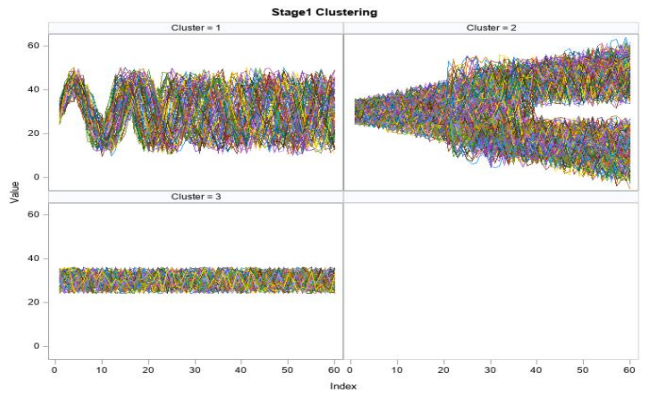

Figure 5: TS Segmentation after Stage_1 Clustering

Figure 6: Segmentation after Stage_2 Clustering

5 PERFORMANCE MEASURE

A. Processing Time

The average processing time is calculated by using formula. Where i represent reduction approach, j represents a number of times run the system for the same dataset and T is total execution time of system.

$$\text{Avg. Processing Time} = \frac{\sum_{i=1}^{5}\sum_{j=1}^{5} T_{i,j}}{5}$$

B. Accuracy

The proposed system result is compared with SAS® Enterprise Miner™ (TS similarity node) using accuracy. Accuracy of the proposed system is calculated using the following equation-

$$\text{Accuracy} = \frac{P}{P+N} * 100$$

Correctly Classified	**P**
Incorrectly Classified	**N**

Table I: Confusion Matrix

Table I shows the confusion matrix used for calculating using accuracy equation, where correctly classified time series will be denoted as P and incorrectly classified time series denoted as N.

Table II shows the average processing time required by the system to complete its task for different data-sets. The average processing time is calculated with each reduction approach by running the system 5 runs.

Table III shows system accuracy with available reduction approaches.

We have measure accuracy with respective of SAS as well as UCI repository datasets. K_1 and K_2 are the numbers of the cluster used at stage_1 and stage_2 clustering respectively. The maximum accuracy of the system lies between 82-100%.

Index	Data-Set	Avg. Processing Time (Sec)
1	Istanbul StockM (UCI)	2.58
2	Snacks Data (SAS)	7.98
3	Synthetic Control Chart (UCI)	10.82
4	Sales Data (UCI)	11.91

Table II: Processing Time

Index	Data-Set	K	SVD	DWT	DFT	LSAS	LSAM
1	Price Data (SAS)	K_1=1 K_2=3	100%	100%	100%	100%	100%
2	Snack Data (SAS)	K_1=1 K_2=3	94%	94%	94%	94%	94%
3	Istanbul StockM (UCI)	K_1=1 K_2=3	78%	100%	89%	89%	89%
4	DowJones StockM (UCI)	K_1=1 K_2=5	100%	100%	100%	100%	100%
5	Synthetic Control Chart (UCI)	K_1=3 K_2=4	82%	81%	82%	82%	82%

Table III: Performance Analysis

6 Conclusion

A. Conclusion

Segments are a group of time series variables that shows similar behavior. Segments are helpful to determine various patterns of time series. Stage_1 clustering based on correlation analysis, which helps to find out TS clusters based on their linear relationships. This staging system extracts segments based on patterns like trend, seasonality, cyclic and normal. Stage_2 clustering based on a reduction approach that reduces time series into a single dimension. This leads to reducing the processing time required for clustering. This stage helps to segment TS data extracted from stage1, based time series magnitude.

B. Future Enhancement

- Moving to Higher Level TS Hierarchy

Currently, the system requires time series data to be in a lower hierarchy. Higher level hierarchy data need to convert into the lower level. In the future, we will be working on a higher level of hierarchical data.

- Cluster Optimization

The current system requires the user to put a number of the cluster for each clustering phase. Cluster optimization will lead to dynamically select clusters based on data.

Acknowledgment

I take this opportunity to express my deep sense of gratitude towards Mr. Sagar Mainkar, Manager, SAS® Research and Development Pune, for giving me this project for dissertation work and his valuable guidance.

References

[1] H. Bunke, A. Kandel, M. Last, "Segmenting Time Series: A Survey And Novel Approach," in Book: Data Mining in Time Series Database, World Scientific Press, 2004, ch. 1, pp. 1-21.

[2] M. Lovric, M. Milanovic, M. Stamenkovicr, "An Algorithmic Method for Segmentation of Time Series: An Overview," Journal of Contemporary Economic and Business Issues (JCEBI), vol. 1, pp. 31-53, 2014.

[3] T. Fu, "A Review on Time Series Data Mining," Engineering Applications of Artificial Intelligence, vol. 24, pp. 164-181, 2011.

[4] "Time Series Analysis and Forecasting," Online Available:https://www.cengage.com/resource_uploads/downloads/0840062389_347257.pdf

[5] "SAS® Time Series Studio 14.1: User's Guide," Online Available:https://support.sas.com/documentation/onlinedoc/forecast/14.1/tsxplug.pdf

[6] S. Aghabozorgi, A. Shirkhorshidi, T. Wah, "Time Series Clustering– A Decade Review," Elsevier: Information Systems, vol. 53, pp. 16-38, 2015.

[7] P. Rodrigues, J. Gama, J. Pedroso, "Hierarchical Clustering of Time-Series Data Streams," IEEE Transactions on Knowledge and Data Engineering, vol. 20, pp. 615-627, 2008.

[8] P. D'Urso, E. Maharaj, "Autocorrelation-based Fuzzy Clustering of Time Series," Fuzzy Sets and Systems, vol. 160, issue 24, pp. 3565-3589, 2009.

[9] M. Riyadi, D. Pratiwi, A. Irawan, K. Fithriasari, "Clustering Stationary and Non-stationary Time Series Based on Auto-correlation Distance of Hierarchical and K-means Algorithms," International Journal of Advances in Intelligent Informatics, vol. 3, pp. 154-160, 2017.

[10] "TSrepr- Time Series Representation in R," Online Available: https://petolau.github.io/TSrepr-time-series-representations/

[11] T. Lee, R. Zhang, Y. Xiao, J. Dean, "Feature Extraction Methods for Time Series Data in SAS® Enterprise Miner™," SAS Institute Inc, pp. 1-14, 2014.

[12] K. Buza, "Time Series Classification and its Applications," International Conference on Web Intelligence, Mining and Semantics, pp. 1–4, 2018.