Time Series representation for clustering using unbalanced Haar wavelet transformation

Sehun Lee, Changryong Baek

Department of Statistics

Sungkyunkwan University

November 15, 2021

불균형 Haar 웨이블릿 변환을 이용한 군집화를 위한 시계열 표현

Context

1. **Introduction**

2. **Method**

3. **Simulation**

4. **Discussion**

Introduction

- Recent time series data tend to be very high-dimensional and high-frequency.
- Due to heavy computation, many studies have been conducted on the dimension reduction method to efficiently handle classification and clustering.

  e.g. DFT(discrete fourier transformation), DWT(discrete wavelet transformation),

  PAA(Piecewise aggregate approximation).

Method

PAA(Piecewise aggregate approximation)(Keogh, 2001).

- PAA is one of the ways to

  reduce time series dimensions.

- Divide into segments of the same size(blocks)

  and average each segment.

- Development: SAX

  (symbolic aggregate approximation)

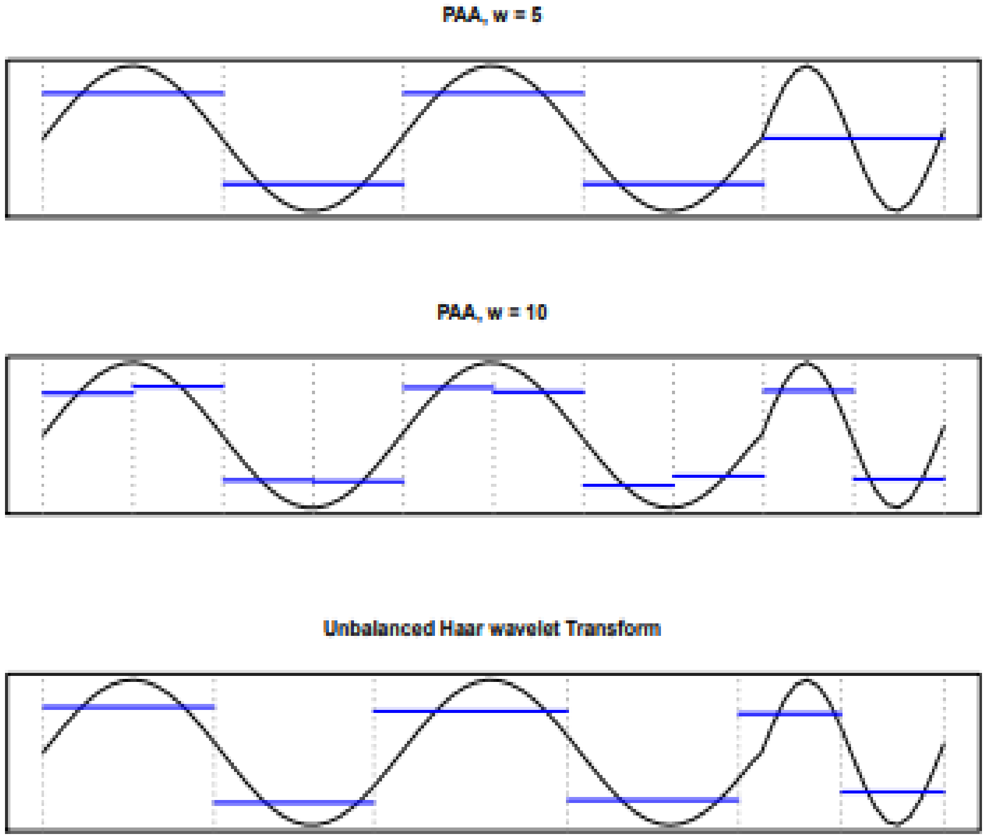


Figure 1.1.

Method

SAX(Symbolic aggregate approximation)(Lin, 2007).

- A method of converting to discrete symbolic data after dimension reduction through PAA.

Features	Limitations
Dimension reduction effect. Apply to discrete data. e.g. text processing, Bioinformatics	Performance depends on the number of segments(w) which determined by the user. (not data adaptive) There is tradeoff according to w. (cf. Figure 1.1.)

Method

DUHT(Discrete Unbalanced Haar Wavelet Transformation)(Fryzlewicz, 2007).

- A method of calculate local mean depending on the data to approximate the time series.

Features
The number of segments is not required and each segment has a different size.
The performance of classification and clustering is improved by **reducing information loss** compared to the existing method.
(Dimension reduction works well). (cf. Figure 1.1.)
Possible to improve the performance of SAX.

Method

SAX Procedure

1. Normalize the time series with different offsets and amplitude.
2. Apply PAA transformation

   2-1. Divide the time series of length n into w equal-sized segments. (local mean)

   $X = \{X_1, \cdots, X_n\}$ to $\bar{X} = \{\overline{X_1}, \cdots, \overline{X_\mathrm{w}}\}$

   2-2. Convert into symbolic data.

   a is the number of symbolic, and find breakpoints β which divide the symbol areas.

$$P(\beta_j \le X < \beta_{j+1}) = \frac{1}{a},\ \ X \sim N(0{,}1)$$

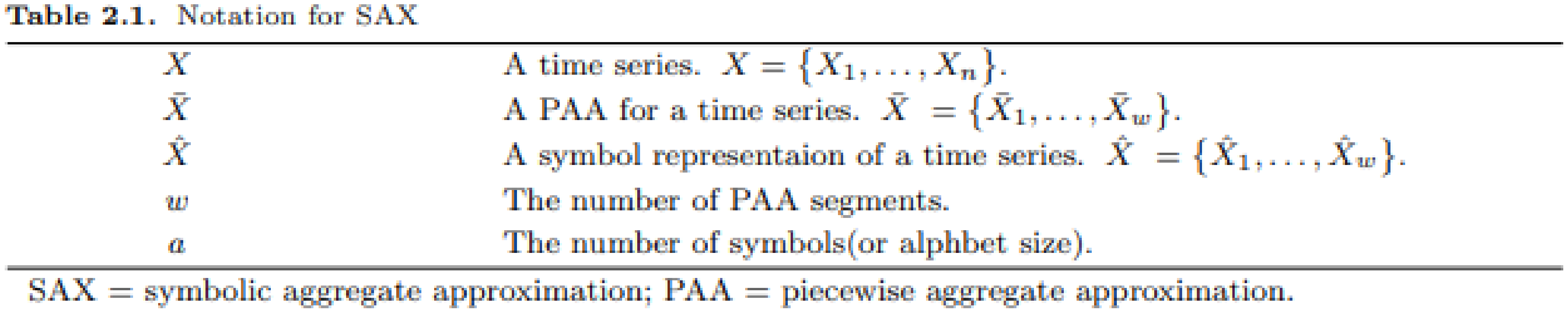


Table 2.1. Notation for SAX

X	A time series. $X = \{X_1, \dots, X_n\}$.
$\bar{X}$	A PAA for a time series. $\bar{X} = \{\bar{X}_1, \dots, \bar{X}_w\}$.
$\hat{X}$	A symbol representaion of a time series. $\hat{X} = \{\hat{X}_1, \dots, \hat{X}_w\}$.
w	The number of PAA segments.
a	The number of symbols(or alphbet size).

SAX = symbolic aggregate approximation; PAA = piecewise aggregate approximation.

Method

3. Let $a_j,\ j = 1, \cdots, a$ symbol for a.

4. If $\beta_j \le \overline{X}_i < \beta_{j+1},\quad \widehat{X}_i = a_j$

The clustering of SAX is calculated by the distance measure MINDIST.

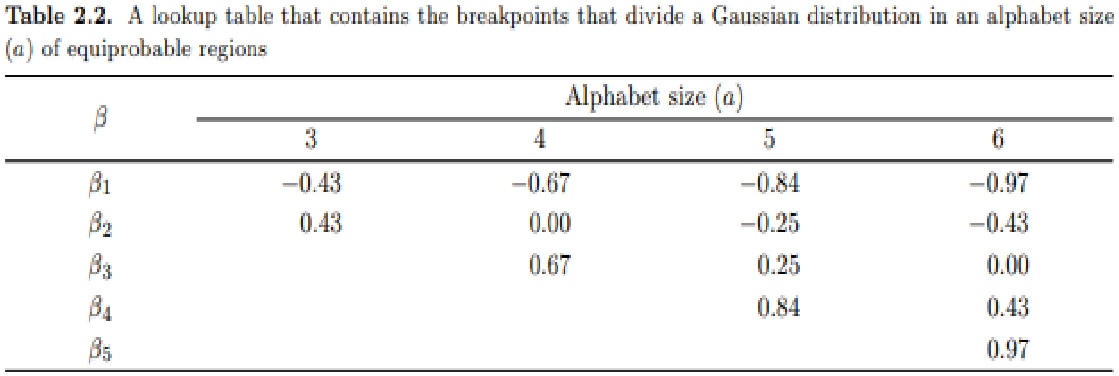


Table 2.2. A lookup table that contains the breakpoints that divide a Gaussian distribution in an alphabet size (a) of equiprobable regions

β	Alphabet size (a)			
	3	4	5	6
β_1	−0.43	−0.67	−0.84	−0.97
β_2	0.43	0.00	−0.25	−0.43
β_3		0.67	0.25	0.00
β_4			0.84	0.43
β_5				0.97

Figure 2.1.

a=3, discrete symbol = {c,a,a,a,b,b,c,b}

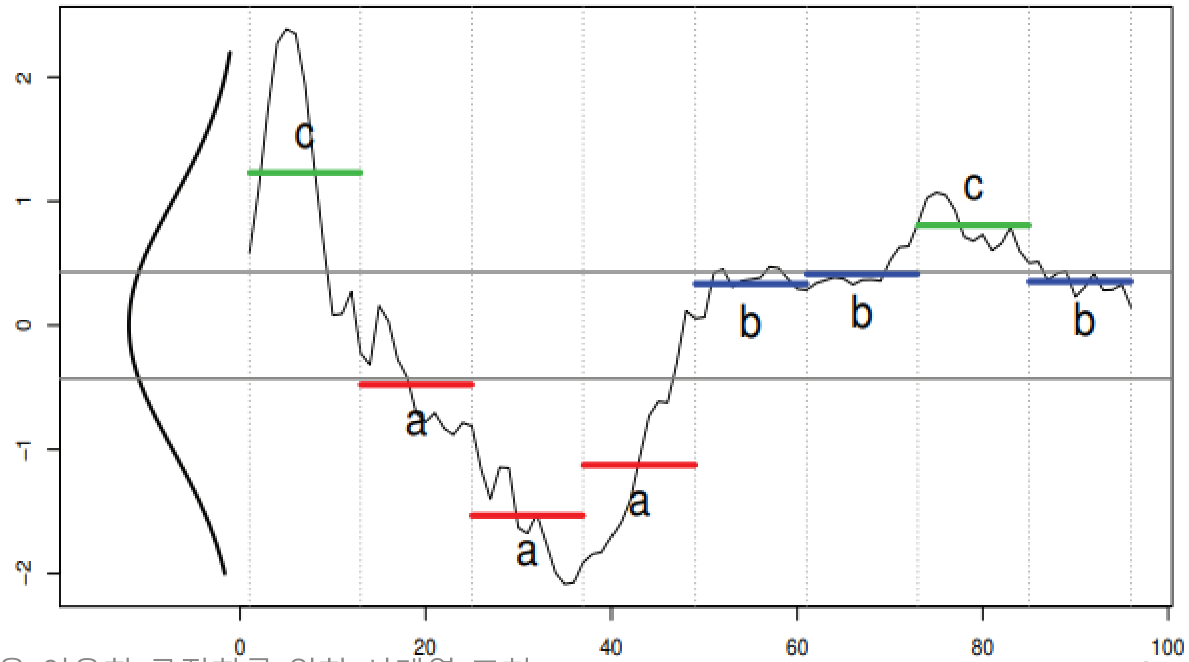


Figure 2.1.

Method

DUHT Procedure

Unbalanced Haar vector: $\psi_{s,b,e}$ (orthnormal basis).

$$\psi_{s,b,e}(l) = \left(\frac{1}{b-s+1} - \frac{1}{e-s+1}\right)^{\frac{1}{2}} 1_{\{s\leq\, l\leq\, b\}} - \left(\frac{1}{e-b} - \frac{1}{e-s+1}\right)^{\frac{1}{2}} 1_{\{b+1\leq\, l\leq\, e\,\}}.$$

Where s: start point(≥1), b: break point , e: end point (≥n)

1. $\mathrm{s}_{0,1} = 1, \mathrm{e}_{0,1} = \ \mathrm{n}$

2. First break point, $b_{\{0,1\}} = argmax_b \left|\left\langle X, \psi_{\{s_{0,1},b,e_{0,1}\ \}}\right\rangle\right|$

3. Then, $\psi^{0,1} = \psi_{\{s_{0,1},b_{0,1},e_{0,1}\}}$

4. Repeat until no more vectors are generated. (generally, $\psi^{j,k} = \psi_{\{s_{j,k},b_{j,k},e_{j,k}\}}$)

Define unbalanced Haar coefficient, $d_{\{j,k\}} = \langle X, \psi^{j,k}\rangle$

Time series X, $X_i = \sum_{j,k} d_{j,k}\psi^{j,k}(i)\ , i = \ 1, \cdots, n$ (orthonormality)

Method

DUHT Procedure

$$\overline{X_{s_{j,k},e_{j,k}}} = \frac{1}{e_{j,k} - s_{j,k} + 1} \sum_{\mathrm{i}=\mathrm{s}_{\mathrm{j},\mathrm{k}}}^{e_{j,k}} \mathrm{X}_{\mathrm{i}} \rightarrow \overline{X_{s_{j,k},e_{j,k}}} = \sum_{\{j\prime<j,k\}} d_{j',k'} \psi^{j',k}(i), i = s_{j,k}, \cdots, e_{j,k}\ (by\ UHT).$$

Why DUHT?

DUHT decomposes time series in finer and finer regimes of changes in local mean level.

This transformation seems to be the one among available methods that leads to simplest structure of inter-arrivals and jumps from zero of the series of changes in mean levels(Baek and Pipiras, 2009).

Method

Small DUHT coefficient, $d_{\{j,k\}}$ can be interpreted as noise.

- Remove small $d_{\{j,k\}}$. by hard thresholding , $d_{\{j,k\}} \rightarrow \tilde{d}_{j,k}$, $X_i \rightarrow \widetilde{X}_i$

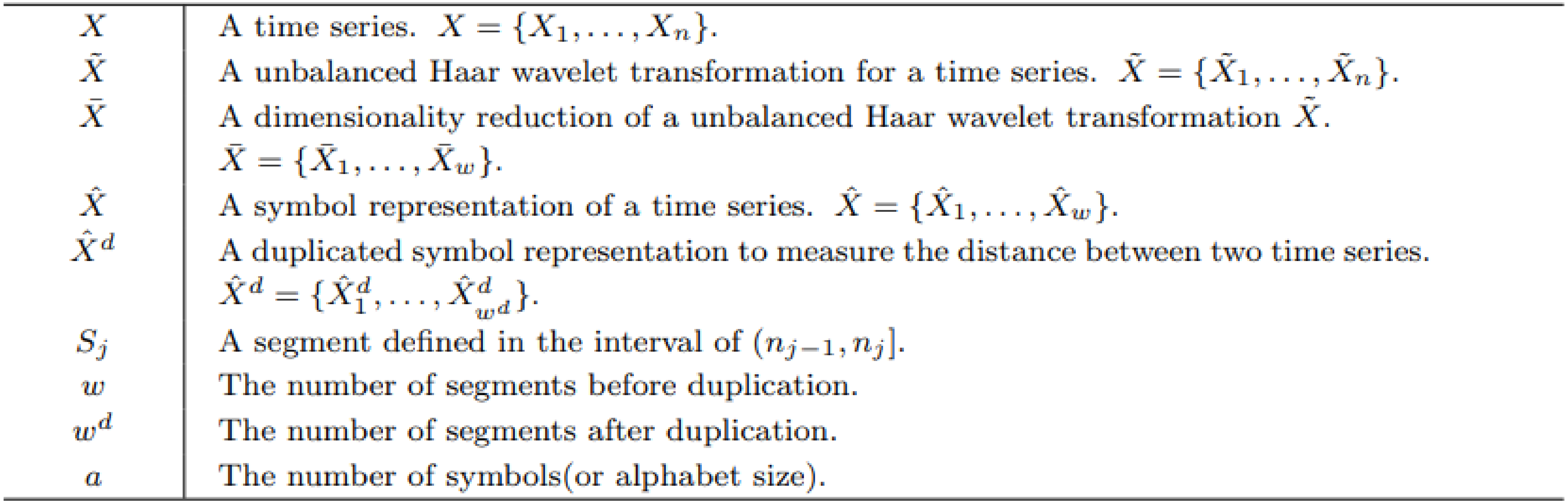
To prevent imbalance,

- Set p, $\max\{\frac{|\psi^{j,k}|^{+}}{|\psi^{j,k}|}, \frac{|\psi^{j,k}|^{-}}{|\psi^{j,k}|}\} \leq p, p \in [\frac{1}{2}, 1)$

Let $|\psi^{j,k}|$, $|\psi^{j,k}|^{+}$ and $|\psi^{j,k}|^{-}$ denote the number of non-zero, positive and negative components of vector $\psi^{j,k}$, respectively.

Method

SAX based on DUHT Procedure



X	A time series. $X = \{X_1, \ldots, X_n\}$.
$\tilde{X}$	A unbalanced Haar wavelet transformation for a time series. $\tilde{X} = \{\tilde{X}_1, \ldots, \tilde{X}_n\}$.
$\bar{X}$	A dimensionality reduction of a unbalanced Haar wavelet transformation $\tilde{X}$. $\bar{X} = \{\bar{X}_1, \ldots, \bar{X}_w\}$.
$\hat{X}$	A symbol representation of a time series. $\hat{X} = \{\hat{X}_1, \ldots, \hat{X}_w\}$.
$\hat{X}^d$	A duplicated symbol representation to measure the distance between two time series. $\hat{X}^d = \{\hat{X}^d_1, \ldots, \hat{X}^d_{w^d}\}$.
S_j	A segment defined in the interval of $(n_{j-1}, n_j]$.
w	The number of segments before duplication.
w^d	The number of segments after duplication.
a	The number of symbols(or alphabet size).

Method

SAX based on DUHT Procedure

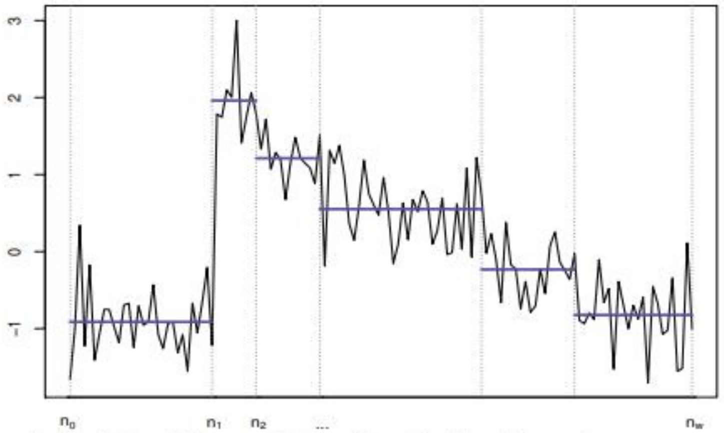


Figure 3.1. Example of unbalanced Haar wavelet transformation for a time series.

Time series points belonging to the same segment will have the same constant value.

$$\widetilde{X_i} = a_j \ \rightarrow \ \overline{\mathrm{X}}\mathrm{i} = a_j, j = 1, \cdots, w.$$

Discretization is performed in the same way as page 8, number 3.

Now we get discrete symbol, $\boldsymbol{\widehat{X} = \{\, \widehat{X_1}, \cdots, \widehat{X_w}\, \}}$**.**

Method

Compare two transformed time series, $\hat{Q} = \{\widehat{Q_1}, \cdots, \widehat{Q_{w_q}}\}$, $\hat{\mathrm{C}} = \{\widehat{\mathrm{C}_1}, \cdots, \widehat{\mathrm{C}_{\mathrm{w_c}}}\}$

1) Different break points

2) Different length.

→ Hard to compute distance.

Solve this problem by define the union set of break points.

e.g. $\hat{Q}$ = {a,b,c,d} , break points = {5, 10, 20}

$\hat{\mathrm{C}}$ = {a,b,c,b,e}, break points = {5, 10, 15, 20}

Find $\widehat{Q^d}$ = {a,b,c,c,d} (duplicate) and compare it with $\widehat{C^d}$.

Simulation

Classification

28 datasets from UCR archives.

Classification error, Figure 4.1.

Compression ratio, Figure 4.2.

(Euclidean :1)

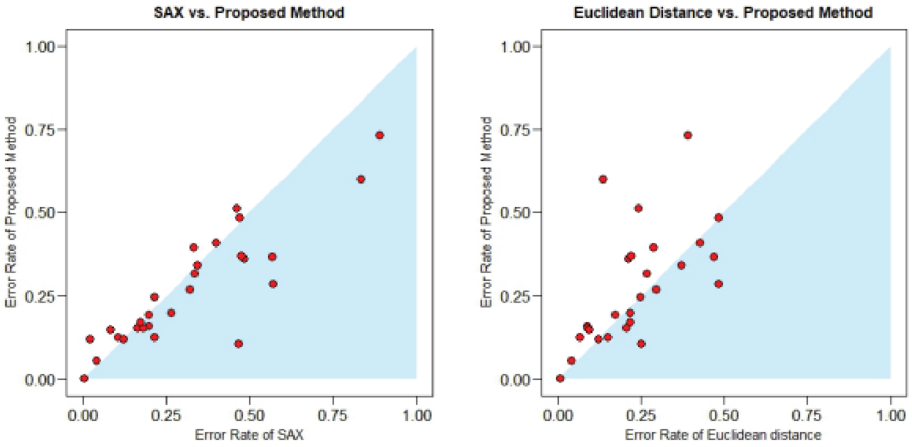


Figure 4.1. Comparision of 1-NN classification error rate on 28 datasets. 1-NN = 1-Nearest Neighbor classification.

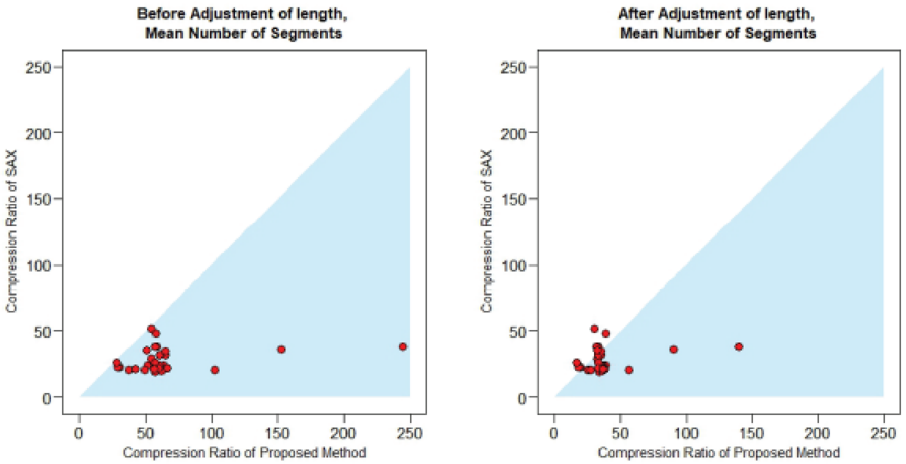


Figure 4.2. Comparision of compression ratio on 28 datasets.

Simulation

Hierarchical Clustering

UCR archives Contro Chart

Level: Normal, cyclic, trend

Method: Proposed Method, SAX, Euclidean distance

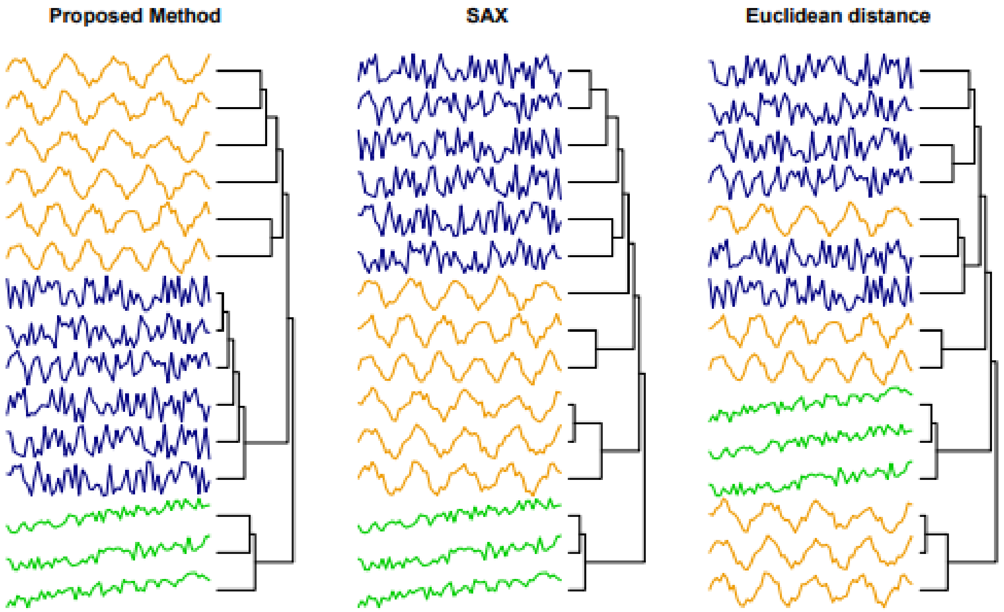


Figure 4.3. Hierarchical clustering of the control chart dataset. SAX = symbolic aggregate approximation.

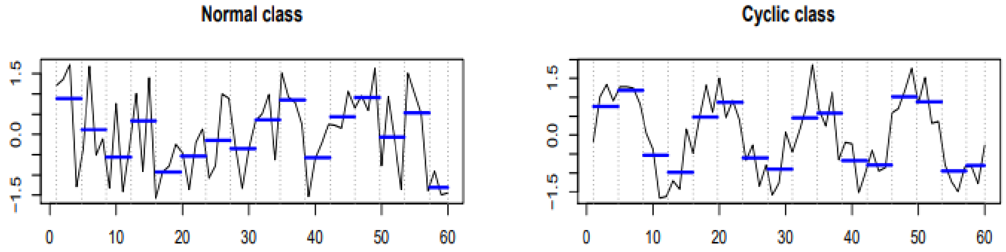


Figure 4.4. Normal and cyclic class converted by the piecewise aggregate approximation.

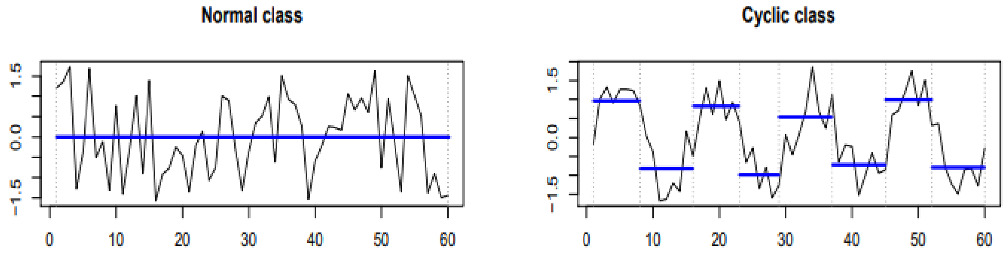


Figure 4.5. Normal and cyclic class converted by the unbalanced Haar wavelet transformation.

Conclusion

Summary

1. DUHT needs more calculation than PAA, but remove the ambiguity of setting w

   and set w depending on the data.

2. Dimensions are effectively reduced

   while preserving important pattern information by dividing into segments of different sizes.

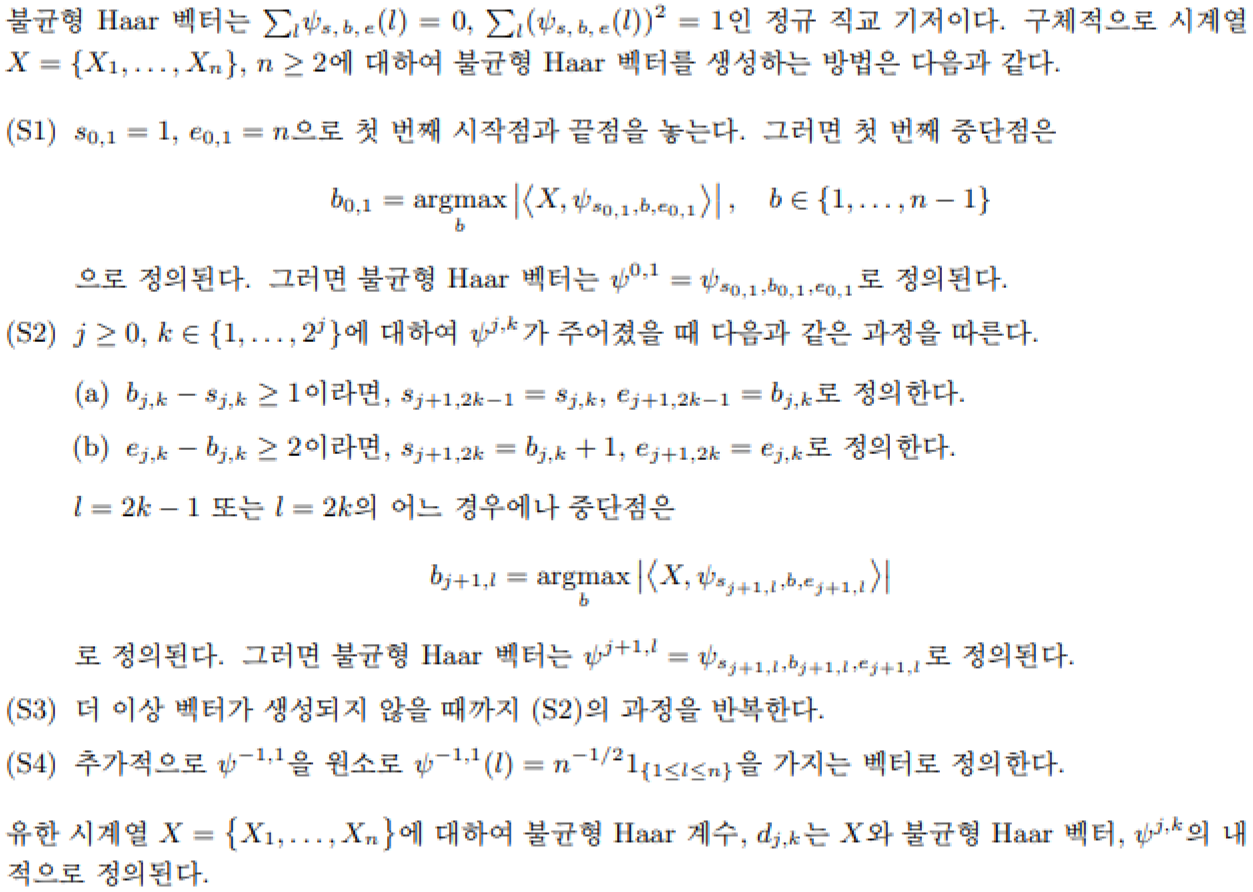
Future study

1. p selection problem. If the break point is too close to the start and end points, the number of data

   is insufficient to find a point of change.

2. a(the number of letters) selection problem.

3. Change thresholding method.



불균형 Haar 벡터는 $\sum_l \psi_{s,\,b,\,e}(l) = 0$, $\sum_l (\psi_{s,\,b,\,e}(l))^2 = 1$인 정규 직교 기저이다. 구체적으로 시계열 $X = \{X_1, \dots, X_n\}$, $n \geq 2$에 대하여 불균형 Haar 벡터를 생성하는 방법은 다음과 같다.

(S1) $s_{0,1} = 1$, $e_{0,1} = n$으로 첫 번째 시작점과 끝점을 놓는다. 그러면 첫 번째 중단점은

$$b_{0,1} = \underset{b}{\operatorname{argmax}} \left|\langle X, \psi_{s_{0,1},b,e_{0,1}}\rangle\right|, \quad b \in \{1, \dots, n-1\}$$

으로 정의된다. 그러면 불균형 Haar 벡터는 $\psi^{0,1} = \psi_{s_{0,1},b_{0,1},e_{0,1}}$로 정의된다.

(S2) $j \geq 0$, $k \in \{1, \dots, 2^j\}$에 대하여 $\psi^{j,k}$가 주어졌을 때 다음과 같은 과정을 따른다.

(a) $b_{j,k} - s_{j,k} \geq 1$이라면, $s_{j+1,2k-1} = s_{j,k}$, $e_{j+1,2k-1} = b_{j,k}$로 정의한다.

(b) $e_{j,k} - b_{j,k} \geq 2$이라면, $s_{j+1,2k} = b_{j,k} + 1$, $e_{j+1,2k} = e_{j,k}$로 정의한다.

$l = 2k - 1$ 또는 $l = 2k$의 어느 경우에나 중단점은

$$b_{j+1,l} = \underset{b}{\operatorname{argmax}} \left|\langle X, \psi_{s_{j+1,l},b,e_{j+1,l}}\rangle\right|$$

로 정의된다. 그러면 불균형 Haar 벡터는 $\psi^{j+1,l} = \psi_{s_{j+1,l},b_{j+1,l},e_{j+1,l}}$로 정의된다.

(S3) 더 이상 벡터가 생성되지 않을 때까지 (S2)의 과정을 반복한다.

(S4) 추가적으로 $\psi^{-1,1}$을 원소로 $\psi^{-1,1}(l) = n^{-1/2}1_{\{1 \leq l \leq n\}}$을 가지는 벡터로 정의한다.

유한 시계열 $X = \{X_1, \dots, X_n\}$에 대하여 불균형 Haar 계수, $d_{j,k}$는 X와 불균형 Haar 벡터, $\psi^{j,k}$의 내적으로 정의된다.

$$d_{j,k} = \langle X,\, \psi^{j,k}\rangle. \tag{2.6}$$