The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.6.707
31(6), 707–719

Time series representation for clustering using unbalanced Haar wavelet transformation

Sehun Lee[a] · Changryong Baek[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received August 8, 2018; Revised September 20, 2018; Accepted November 15, 2018)

Abstract

Various time series representation methods have been proposed for efficient time series clustering and classification. Lin *et al.* (*DMKD*, **15**, 107–144, 2007) proposed a symbolic aggregate approximation (SAX) method based on symbolic representations after approximating the original time series using piecewise local mean. The performance of SAX therefore depends heavily on how well the piecewise local averages approximate original time series features. SAX equally divides the entire series into an arbitrary number of segments; however, it is not sufficient to capture key features from complex, large-scale time series data. Therefore, this paper considers data-adaptive local constant approximation of the time series using the unbalanced Haar wavelet transformation. The proposed method is shown to outperforms SAX in many real-world data applications.

Keywords: time series representation, SAX, unbalanced Haar wavelet transformation, classification, clustering

1. 서론

본 연구는 시계열 자료의 분류와 군집화에 대해서 효과적인 표현 방법을 연구한다. 최근 기술 및 저장 장치의 발전은 시계열 자료를 매우 고차원(high dimensionality) 및 고빈도(high frequency)인 특징을 지닌 자료로 만들었다. 이런 풍부한 자료는 정보량의 증가에 긍정적인 영향을 미쳤지만 이런 정보를 학습하여 처리하기 위해 필요한 계산량은 기하 급수적으로 증가하였다. 따라서 지난 세대 동안 시계열 분류와 군집화의 분석을 효율적으로 처리하기위한 자료의 요약, 즉 차원의 축소 방법에 대해서 많은 연구가 진행되었다. 대표적인 연구로는 이산 푸리에 변환(discrete Fourier transform; DFT) (Faloutsos 등, 1994), 이산 웨이블릿 변환(discrete wavelet transform; DWT) (Chan과 Fu, 1999), piecewise aggregate approximation (PAA) (Keogh 등, 2001) 등을 찾아 볼 수 있으며 전반적인 시계열 군집화에 대해서는 Aghabozorgi 등 (2015)를 참조하기 바란다.

This work was supported in part by the Basic Science Research Program from the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1 A1A1A05000831).

[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: crbaek@skku.edu

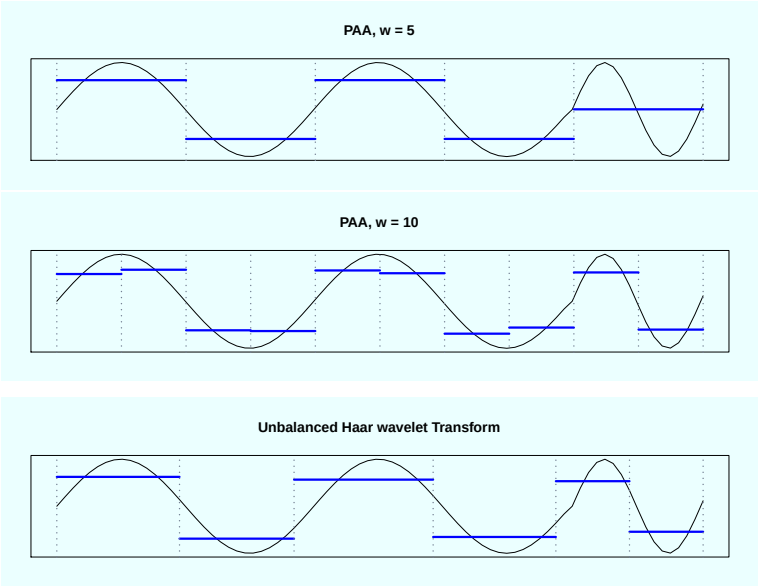

Figure 1.1. Example of and unbalanced Haar wavelet transform for a time series. PAA = piecewise aggregate approximation.

이 중 Keogh 등 (2001)이 제안한 PAA는 시계열을 동일한 크기의 세그먼트(블록)로 나눈 후 각각의 세그먼트를 평균함으로써 시계열 데이터의 차원을 축소하는 방법으로 국소 평균을 이용한 시계열 자료의 근사라고 생각 할 수 있다. 이 방법을 기반으로 Lin 등 (2007)은 PAA를 이용하여 시계열의 차원을 축소한 후 이산 심볼릭 자료로 변환하는 symbolic aggregate approximation (SAX) 방법을 제안하였다. SAX 방법은 차원을 효과적으로 줄일 뿐만 아니라 SAX 표현으로 변환된 시계열은 해싱(hashing)과 서픽스 트리(suffix tree) 등 이산 데이터에 대해 정의되는 텍스트 처리(text processing)와 생물정보학으로부터의 데이터 구조 및 알고리즘을 활용할 수 있는 장점이 있다.

하지만 SAX는 몇 가지 한계점을 가진다. SAX의 성능은 세그먼트의 수 w(혹은 블럭의 개수)에 의존하지만, w의 값은 사용자에 의해 임의적으로 결정되며 데이터에 의존(data adaptive) 하지 않기 때문에 w의 크기에 따라 성능이 현저히 달라질 수 있다. 대부분의 커널을 이용한 함수 추정 방법이 가지고 있는 편의-분산의 이율 배반성(bias-variance trade-off)을 SAX에서도 발견할 수 있다. 즉, 원 시계열에 대한 정보의 손실을 최소화하기 위해서는 w의 값을 크게 해야 하지만, 이는 차원 축소의 효과를 줄인다. 예를 들어 Figure 1.1은 w의 값에 따른 PAA의 예를 보여준다. w의 값이 5일 경우, PAA는 마지막 세그먼트의 영역에 속하는 시계열의 패턴을 올바르게 표현하지 못하며 w의 값을 두 배인, 10으로 설정하였을 때 보다 시계열의 패턴을 올바르게 표현하고 있음을 관찰할 수 있다. 하지만 자료의 축소 및 분산의 증가 관점에서는 무한정 w를 늘릴 수 없어 최적의 w를 찾는 것은 SAX 방법을 적용하는 데 있어 매우 중요하다.

본 연구는 Fryzlewicz (2007)에 의해 제안된 이산 불균형 Haar 웨이블릿 변환(discrete unbalanced Haar wavelet transformation; DUHT)을 이용하여 국소 평균 수준의 변화를 자료에 의존적(data dependent)으로 계산하여 시계열을 근사하는 방법에 대해서 제안한다. 예를 들어, Figure 1.1의 하단은 PAA의 예와 동일한 시계열에 대해 불균형 Haar 웨이블릿 변환을 적용한 결과로, 시계열은 6개의 동일

Table 2.1. Notation for SAX

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

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

한 국소 평균 수준을 가지는 시계열로 근사되며 근사된 시계열은 원 시계열의 패턴을 올바르게 표현하는 것을 확인할 수 있다. 이와 동시에 DUHT의 경우 각 세그먼트의 크기가 다르기 때문에 효과적인 차원의 축소도 가능하다. 따라서, 이는 PAA 대신 불균형 Haar 웨이블릿 변환을 이용하여 차원을 축소함으로써 SAX의 한계점을 보완할 수 있음을 보여준다. 보다 구체적으로 본 논문에서는 기존 SAX에서 PAA 대신 불균형 Haar 웨이블릿 변환을 이용하여 시계열의 차원을 축소하는 시계열 표현 방법을 제안한다. 제안한 방법은 기존 SAX와 달리 모수로서 세그먼트의 수를 요구하지 않으며 시계열의 국소 평균 수준의 변화에 따라 시계열의 차원을 축소하므로 기존 방법보다 원 시계열에 대한 정보의 손실을 줄여 분류와 군집화의 성능을 향상시킬 수 있다.

본 논문의 구성은 다음과 같다. 제 2장에서는 SAX와 불균형 Haar 웨이블릿 변환에 대해 소개한다. 제 3장에서는 불균형 Haar 웨이블릿 변환을 이용하여 차원을 축소하는 수정된 SAX 방법을 제안하고, 제 4장에서는 최근접 이웃 분류(1-Nearest Neighbor classification; 1-NN)와 계층적 군집화(hierarchical clustering)를 통해 기존 방법과 본 논문에서 제안한 방법을 비교하고 그 성능을 검증한다. 마지막으로 제 5장에서는 본 논문에서 제안한 방법의 장단점과 논의점에 대해 이야기하고 마무리한다.

2. 배경 연구

본 장에서는 본 논문의 핵심 방법인 SAX 및 DUHT에 대해서 간략히 소개하고 수식 표현 (notation)을 정리한다.

2.1. SAX

Lin 등 (2007)에 의해 제안된 SAX은 PAA (Keogh 등, 2001)를 통해 차원을 축소한 후 이산 심볼로 변환하는 시계열 표현 방법이다. Table 2.1은 SAX을 설명하는데 사용되는 표기법을 요약한 것이다. 우선 서로 다른 변위(offsets)와 진폭(amplitudes)을 가진 시계열을 비교하기 위해 정규화(normalization)한 후 PAA로 변환한다. PAA에 의해 길이 n의 시계열 $X = \{X_1, \ldots, X_n\}$는 w개의 동일한 크기의 세그먼트로 나누어지고 각각의 세그먼트의 평균인, w차원의 벡터 $\bar{X} = \{\bar{X}_1, \ldots, \bar{X}_w\}$로 변환된다. 즉, $\bar{X}$의 i번째 요소는 다음의 식에 의해 계산된다.

$$\bar{X}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} X_j.$$

국소 평균으로 축소된 시계열 자료는 다음의 과정을 통해 심볼릭자료로 변환된다. 먼저 정규 분포를 a개의 동일한 확률 영역으로 나누어, 심볼 영역을 나누는 구분점(breakpoints) $\beta_1, \ldots, \beta_{a-1}$를 찾는다. 이는 수식으로

$$P(\beta_j \leq X < \beta_{j+1}) = \frac{1}{a}, \quad X \sim N(0,1) \tag{2.1}$$

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

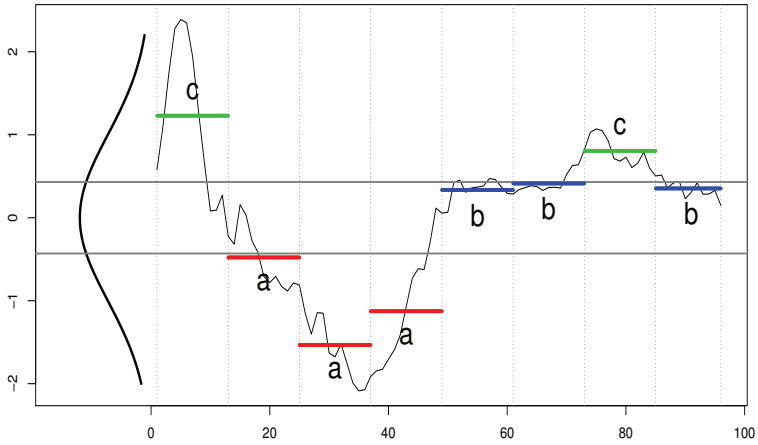

Figure 2.1. Example of symbolic aggregate approximation for a time series.

을 만족하는 값으로 정의되며 $\beta_0 = -\infty$, $\beta_a = \infty$이다. 예를 들어, Table 2.2는 심볼의 개수(a)에 따른 구분점의 값을 보여준다.

a개의 구간에 대해서 심볼을 α_j, $j = 1, \ldots, a$라 정의 하면 PAA로 변환된 시계열은 $\bar{X}_i$가 속한 구간의 심볼로 치환된다. 만약 $\beta_{j-1} \leq \bar{X}_i < \beta_j$이라면

$$\hat{X}_i = \alpha_j \tag{2.2}$$

알파벳으로 치환된다. Figure 2.1은 SAX의 예로, 길이 96의 시계열을 세그먼트 수 $w = 8$, 알파벳 크기 $a = 3$으로 SAX에 의해 변환한다면 8개의 문자(string)을 갖는 이산 심볼, $\{c, a, a, a, b, b, c, b\}$로 변환된다.

SAX 변환된 두 시계열 Q와 C는 다음과 같이 정의된 거리측도인 MINDIST에 의해 거리를 계산하고 이를 기반으로 군집화가 진행된다.

$$\text{MINDIST}\left(\hat{Q}, \hat{C}\right) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} \left(\text{dist}\left(\hat{Q}_i, \hat{C}_i\right)\right)^2}. \tag{2.3}$$

여기서 dist 함수는 다음의 식에 의해 정의된다.

$$\text{dist}(\alpha_i, \alpha_j) = \begin{cases} 0, & \text{if } |i - j| \leq 1, \\ \beta_{\max(i,j)-1} - \beta_{\min(i,\,j)}, & \text{otherwise.} \end{cases} \tag{2.4}$$

Table 2.3. A lookup table for the MINDIST function when $a = 4$

	α_1	α_2	α_3	α_4
α_1	0	0	0.67	1.34
α_2	0	0	0	0.67
α_3	0.67	0	0	0
α_4	1.34	0.67	0	0

MINDIST의 계산은 Table 2.3과 같이 dist 함수의 값에 대한 테이블을 만들어 조회함으로써 빠르게 계산할 수 있다.

2.2. 불균형 Haar 웨이블릿 변환

전통적인 Haar 웨이블릿 변환은 데이터를 이분 분해(dyadic decomposition)하기 때문에 길이가 2의 거듭제곱인 시계열에 대해서만 정의된다는 단점이 있다. Fryzlewicz (2007)이 사용한 불균형 Haar 웨이블릿 변환은 불균형 Haar 벡터 $\psi_{s,\,b,\,e}$에 기반을 둔다. 불균형 Haar 벡터의 원소 $\psi_{s,\,b,\,e}(l)$는 시계열 자료의 시작점 s, 변화점 b, 끝점 e에 대해서 다음과 같이 정의된다.

$$\psi_{s,\,b,\,e}(l) = \left(\frac{1}{b-s+1} - \frac{1}{e-s+1}\right)^{\frac{1}{2}} 1_{\{s\le l\le b\}} - \left(\frac{1}{e-b} - \frac{1}{e-s+1}\right)^{\frac{1}{2}} 1_{\{b+1\le l\le e\}}. \tag{2.5}$$

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 $X = \{X_1, \ldots, X_n\}$, $n \ge 2$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

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

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

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

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

(b) $e_{j,k} - b_{j,k} \ge 2$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

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

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

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

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

그러면 시계열 X는

$$X_i = \sum_{j,k} d_{j,k}\psi^{j,k}(i), \quad i = 1, \ldots, n \tag{2.7}$$

로 표현된다. 또한 $\bar{X}_{s_{j,k},e_{j,k}} = 1/(e_{j,k} - s_{j,k} + 1)\sum_{i=s_{j,k}}^{e_{j,k}} X_i$라 하면, 이는 불균형 Haar 변환을 통해

$$\bar{X}_{s_{j,k},e_{j,k}} = \sum_{j'<j}\sum_{k'} d_{j',k'}\psi^{j',k'}(i), \quad i = s_{j,k}, \ldots, e_{j,k} \tag{2.8}$$

로 나타낼 수 있다. 따라서 식 (2.5)와 식 (2.8)의 관점에서, 불균형 Haar 변환은 다양한 스케일, j에서 국소 평균 수준의 변화에 따라 시계열을 분해하는 과정이라 할 수 있다 (Baek과 Pipiras, 2009).

불균형 Haar 계수의 작은 값은 국소 평균의 작은 변화, 즉 노이즈로 해석될 수 있다. 따라서 시계열의 국소 평균 수준의 변화를 잘 설명하도록 시계열을 근사하기 위해선 노이즈로 해석될 수 있는 불균형 Haar 계수의 작은 값들을 제거해야 한다. 다음의 식과 같이 하드 임계처리(hard threshoding)를 이용하여 작은 불균형 Haar 계수를 제거한다.

$$\tilde{d}_{j,k} = d_{j,k} I(\left|d_{j,k}\right| > \lambda). \tag{2.9}$$

여기서 임계치, λ는 전체 임계치(universal threshold)인

$$\lambda = \sigma\sqrt{2\log n} \tag{2.10}$$

이다. 노이즈가 제거된 시계열은 식 (2.7)에서 $d_{j,k}$를 $\tilde{d}_{j,k}$로 대체함으로써 다음과 같이 근사된다.

$$\tilde{X}_i = \sum_{j,k} \tilde{d}_{j,k}\psi_{j,k}(i), \quad i = 1, \ldots, n. \tag{2.11}$$

추가적으로, 불균형 Haar 변환의 계산 복잡도를 줄이기 위해 다음의 가정을 추가한다. 모든 n에 대하여, $j \geq 0$, k에 관하여 일률적으로

$$\max\left\{\frac{b_{j,k} - s_{j,k} + 1}{e_{j,k} - s_{j,k} + 1}, \frac{e_{j,k} - b_{j,k}}{e_{j,k} - s_{j,k} + 1}\right\} \leq p \tag{2.12}$$

가 되도록 하는 고정된 상수 $p \in [1/2, 1)$이 존재한다고 가정한다. 이 조건이 만족되면 불균형 Haar 변환의 계산 복잡도는 $O(n\log n)$이 된다.

3. DUHT를 0용한 SAX

본 절에서는 본 논문에서 제안하고자하는 방법인, 불균형 Haar 웨이블릿 변환을 이용하여 차원을 축소하는 수정된 SAX 방법에 대해서 소개한다.

3.1. 제안한 방법

Table 3.1은 제안한 방법을 설명하는데 사용되는 표기법을 요약한 것이다. 우선 정규화된 시계열 $X = \{X_1, \ldots, X_n\}$가 주어졌을 때, 주어진 시계열을 불균형 Haar 웨이블릿 변환을 이용하여 다음과 같이 변환한다.

$$\tilde{X}_i = \sum_{j,k} \tilde{d}_{j,k}\psi_{j,k}(i), \quad i = 1, \ldots, n.$$

Table 3.1. Notation for the proposed method

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).

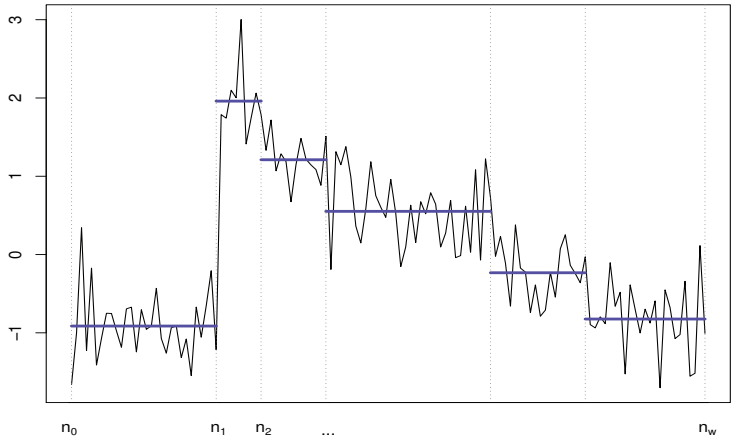

Figure 3.1. Example of unbalanced Haar wavelet transformation for a time series.

여기서 $\tilde{d}_{j,k}$는 식 (2.9)와 식 (2.10)을 통해 임계처리된 것으로, 이때 식 (2.10)의 표준편차 σ는 Donoho와 Johnstone (1994)에 의해 제안된 방법인 불균형 Haar 계수, $d_{j,k}$의 중앙값 절대 편차(median absolute deviation; MAD)로 추정한다. 불균형 Haar 웨이블릿 변환을 통해 변환된 시계열, $\tilde{X}$은 Figure 3.1의 예처럼 동일한 국소 평균을 가지는 w개의 세그먼트로 나누어진다. 따라서 동일한 세그먼트에 속하는 시계열 포인트, $\tilde{X}_i$은 모두 동일한 상수 값을 가지므로, 각 세그먼트 당 하나의 값만을 사용하여 n 차원의 시계열을 w 차원으로 축소한다. 다시말해, 구간 $(n_{j-1}, n_j]$에서 정의되는 세그먼트를 S_j, $j = 1, \ldots, w$라 하면(n_0와 n_w는 각각 0과 n으로 정의한다), S_j에 속하는 데이터 포인트들은 다음과 같이 동일한 상수, a_j를 갖는다.

$$\tilde{X}_i = a_j, \quad \tilde{X}_i \in S_j. \tag{3.1}$$

그러면 축소된 시계열, $\bar{X} = \{\bar{X}_1, \ldots, \bar{X}_w\}$은 다음과 같이 정의된다.

$$\bar{X}_j = a_j, \quad j = 1, \ldots, w. \tag{3.2}$$

축소된 시계열은 기존의 SAX와 동일한 방법으로 식 (2.1)과 식 (2.2)를 이용하여 이산화된다. 최종적으로 n 차원의 시계열, $X = \{X_1, \ldots, X_n\}$는 제안한 방법에 의해 w 차원의 이산 심볼, $\hat{X} = \{\hat{X}_1, \ldots, \hat{X}_w\}$로 변환된다.

3.2. 거리의 계산

제안한 방법에 의해 변환된 두 시계열, $\hat{Q} = \{\hat{Q}_1, \ldots, \hat{Q}_{w_q}\}$와 $\hat{C} = \{\hat{C}_1, \ldots, \hat{C}_{w_c}\}$은 각각 서로 다른 구분점 및 길이, w_q와 w_c를 가지므로 두 시계열의 거리를 계산하기 위해서 두 시계열이 서로 동일한 구분점과 길이, w^d를 가지도록 만들어야한다. 이는 각 세그먼트를 구분해주는 두 시계열의 구분점의 집합의 합집합을 $\{n_0 = 0, n_1, \ldots, n_{w^d} = n\}$을 정의함으로써 해결할 수 있다. 시계열의 전체 구간 $(0, n]$을 두 시계열의 구분점의 합집합으로부터 분할한 구간을 S_j, $j = 1, \ldots, w^d$라 하자. 즉, 세그먼트 S_j는 구간 $(n_{j-1}, n_j]$에서 정의된다. 그러면 다음의 정의된 식

$$\hat{Q}_j^d = \hat{Q}_i, \quad \text{iff } \hat{Q}_i \in S_j,\ j = 1, \ldots, w^d \tag{3.3}$$

에 의해 심볼을 복제(duplicate) 함으로써 $\hat{Q}$은 $\hat{Q}^d = \{\hat{Q}_1^d, \ldots, \hat{Q}_{w^d}^d\}$로 길이가 조정된다. 다른 시계열 $\hat{C}$에 대해서도 동일하게 적용하면 두 시계열은 공통의 구분점을 썼기에 동일한 크기를 가지는 심볼릭 문자열이 된다. 길이가 조정된 두 시계열은 기존 SAX의 거리 측도인 식 (2.3)에서 가중치 부분을 수정한, 다음의 식에 의해 계산되며

$$\text{DISTANCE}\left(\hat{Q}, \hat{C}\right) = \sqrt{\sum_{i=1}^{w^d} (n_i - n_{i-1}) \left(\text{dist}\left(\hat{Q}_i^d, \hat{C}_i^d\right)\right)^2} \tag{3.4}$$

dist 함수는 기존 SAX과 동일한 수식 (2.4)이다. 예를 들어, 원 시계열의 길이가 25일 때, $\hat{Q} = \{a, b, c, d\}$은 구분점 $\{5, 10, 20\}$을 통해서 $\hat{C} = \{a, b, c, b, e\}$은 구분점 $\{5, 10, 15, 20\}$을 통해서 얻어졌다고 가정하자. 그렇다면 $\hat{Q}$와 $\hat{C}$는 각각 다른 구분점을 가지고 크기도 달라 거리를 구할 수 없다. 이를 수정하기 위해서 구분점의 합집합인 $\{5, 10, 15, 20\}$에 대해서 이산 심볼을 다시 찾으면 된다. 즉 $\hat{Q}$에 대해서는 10과 20 사이에 15라는 구분점이 더 있으므로 심볼을 복제하여 최종 심볼은 $\hat{Q}^d = \{a, b, c, c, d\}$가 된다. 이는 $\hat{C}$가 가지고 있는 구분점과도 일치하고 서로 동일한 길이가 되어 수식 (3.4)에 의해 거리를 구할 수 있다.

4. 경험적 평가

4.1. 최근접 이웃 분류

UCR archives로부터 28개의 데이터셋(http://www.cs.ucr.edu/∼eamonn/time_series_data/)을 이용하여 제안한 방법과 기존의 SAX 방법, 원 시계열에 대해 유클리디안 거리(Euclidean distance)을 적용한 방법에 대해 최근접 이웃 분류를 수행하였다. SAX와 유클리디안 거리의 결과는 Lin 등 (2007)에서의 결과를 사용하였다. 기존 방법과의 비교를 위하여 제안한 방법도 Lin 등 (2007)에서와 동일하게 a를 10으로 설정하였다. Table 4.1은 최근접 이웃 분류의 결과를 보여준다. 여기에서 SAX w는 SAX에 의해 변환된 시계열의 평균 세그먼트 수를 나타내고, mean w와 mean w^d는 각각 제안한 방법에 의해 변환된 시계열의 평균 세그먼트 수와 거리 계산을 위해 길이가 조정된 시계열 표현의 평균 세그먼트 수를 나타낸다. EU error와 SAX error, Proposed method error는 각각 유클리디안 거리에 대한 오분류율과 SAX에 대한 오분류율, 제안한 방법에 대한 오분류율을 나타낸다. 각 데이터셋에 대한 더 자세한 정보는 위에서 언급한 UCR archives 사이트에서 확인가능하다.

Figure 4.1은 Table 4.1의 결과를 요약해 보여준다. 그림의 아래쪽 삼각형 영역에 그려진 점들은 제안한 방법이 비교되는 방법보다 오분류율이 더 낮다는 것을 뜻한다. Figure 4.2는 제안한 방법과 SAX 방법의 압축률을 비교한 그림이다. Figure 4.1와 마찬가지로 아래쪽 삼각형 영역에 그려진 점들은 제안한

Table 4.1. Comparision of 1-NN classification error rate on 28 datasets

Dataname	Time series length	SAX w	Mean w	Mean w^d	EU error	SAX error	Proposed method error
Synthetic Control	60	16	4	6	0.120	0.020	0.120
Gun Point	150	64	47	74	0.087	0.180	0.153
CBF	128	32	5	9	0.148	0.104	0.126
FaceAll	131	64	20	35	0.286	0.330	0.395
OSULeaf	427	128	63	116	0.483	0.467	0.483
SwedishLeaf	128	32	21	37	0.211	0.483	0.362
50Words	270	128	53	94	0.369	0.341	0.343
Trace	275	128	26	47	0.240	0.460	0.510
MiddlePhalanx OutlineAgeGroup	80	32	15	23	0.481	0.568	0.285
ProximalPhalanx OutlineAgeGroup	80	32	13	20	0.215	0.263	0.200
ProximalPhalanx TW	80	16	13	20	0.293	0.320	0.270
Two Patterns	128	32	22	39	0.093	0.081	0.150
MALLAT	1024	512	172	291	0.086	0.197	0.159
Wafer	152	64	51	81	0.0045	0.0034	0.0031
DiatomSize Reduction	345	64	61	108	0.065	0.212	0.128
FaceFour	350	128	120	196	0.216	0.170	0.170
Lighting2	637	256	96	168	0.246	0.213	0.246
Lighting7	319	128	51	88	0.425	0.397	0.411
ECG200	96	32	17	29	0.120	0.120	0.120
ECGFiveDays	136	64	35	51	0.203	0.161	0.156
Adiac	176	64	30	51	0.389	0.890	0.731
Yoga	426	128	68	124	0.170	0.195	0.194
Fish	463	128	69	126	0.217	0.474	0.371
Plane	144	64	21	37	0.038	0.038	0.057
Car	577	256	93	169	0.267	0.333	0.317
Beef	470	128	89	140	0.467	0.567	0.367
Coffee	286	128	49	73	0.250	0.464	0.107
OliveOil	570	256	130	149	0.133	0.833	0.600

1-NN = 1-Nearest Neighbor classification; SAX = symbolic aggregate approximation.

방법이 SAX보다 압축률이 높다는 것을, 즉 차원의 축소효과가 큼을 뜻한다. 압축률을 비교하자면 원 시계열은 실수값으로 각 시점마다 32비트(bit)의 메모리 를 필요로 하는 반면, 제안한 방법과 SAX 방법에 의해 이산 심볼로 변환된 시계열에 대해서는 각 심볼 당 $[\log_2 a]$비트만을 필요로 한다. 따라서 압축률은

$$\frac{n \times 32}{w \times \log_2 a}$$

로 계산되어진다. 유클리디안 거리의 방법의 경우 압축률은 1이다. 28개의 데이터셋 중 16개의 데이터셋에서 제안한 방법이 기존 SAX보다 오분류율이 낮고, 10개의 데이터셋에서는 기존 방법보다 높은 오분류율을 보이지만 그 차이가 미미한 것을 확인할 수 있다. 거리 계산을 위해 시계열의 길이가 조정되기 전의 평균 세그먼트의 수에 대한 압축률은 모든 데이터셋에서 기존 방법보다 제안한 방법이 압축률이 높다. 거리 계산을 위해 시계열의 길이가 조정된 후의 평균 세그먼트의 수에 대한 압축률은 8개의 데이터셋을 제외하고 제안한 방법이 기존 방법보다 압축률이 높은 것을 확인할 수 있다. 유클리디안 거리와 비

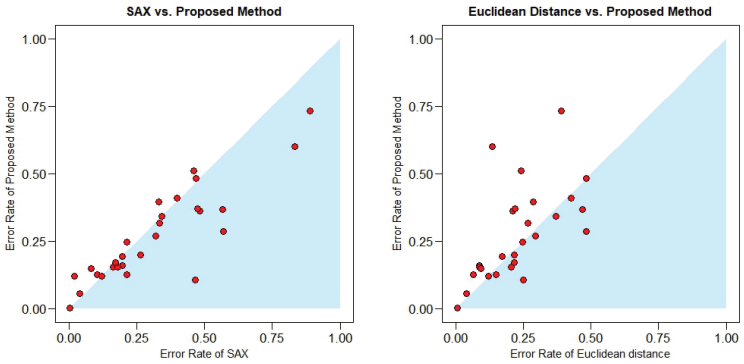

Figure 4.1. Comparision of 1-NN classification error rate on 28 datasets. 1-NN = 1-Nearest Neighbor classification.

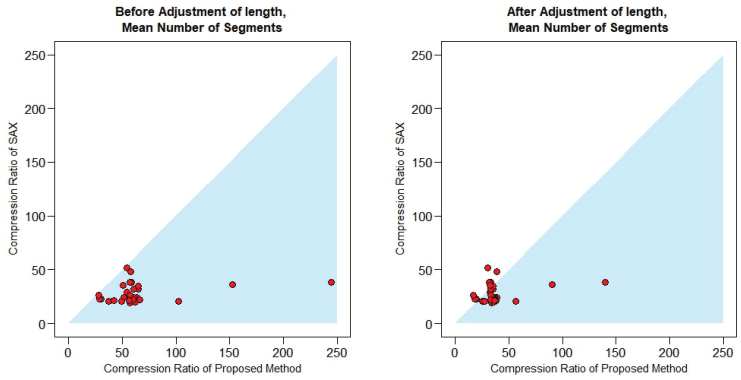

Figure 4.2. Comparision of compression ratio on 28 datasets.

교하였을 때, 제안한 방법이 다소 높은 오분률율을 보이지만 유클리디안 방법은 자료의 압축이 전혀 없으며 제안한 방법의 결과는 심볼의 크기, a를 10으로 고정한 결과로, a의 값을 조정함으로써 오분류율을 낮출 수 있을 것이라 기대한다. 또한 제안한 방법은 시계열의 차원을 축소하고 이산 심볼로 변환함으로써 유클리디안 거리의 방법보다 분류를 더 효율적으로 수행할 수 있다.

4.2. 계층적 군집화

군집화의 성능을 평가하기 위해 라벨의 정보를 가지고 있는 UCR archives로부터의 Control Chart 자료에 대해서 계층적 군집화(hierarchical clustering)을 수행하였다. Figure 4.3은 Control Chart 자료로부터 노말과 순환, 상승추세의 세 가지 계급에 대해서 계층적 군집화를 수행한 결과를 보여준다. 유클리디안 거리와 SAX은 노말과 순환 계급을 올바르게 군집화하지 못하는 것을 볼 수 있다. 반면, 제안한 방법은 모든 계급을 올바르게 군집화한다. 이는 PAA를 이용한 차원 축소가 효과적으로 원 시계열의 특징을 반영하지 못하기 때문으로 Figure 4.4와 Figure 4.5를 비교함으로써 보다 명확히 살펴볼 수 있다. 우리가 제안한 불균형 Haar 웨이블릿 변환은 자료의 형태 및 특성에 따라 변화점을 찾기에 노말 계급과 순환 계급의 패턴을 올바르게 표현하는 것을 확인할 수 있다.

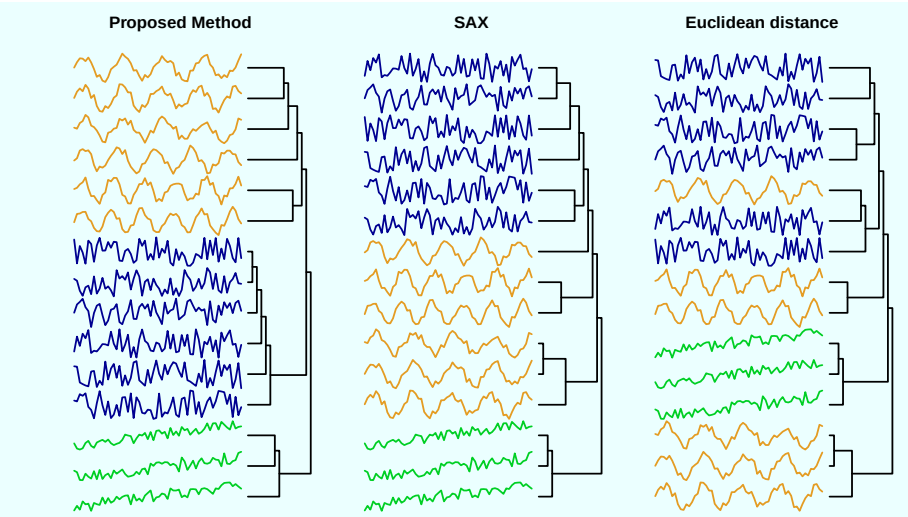

Figure 4.3. Hierarchical clustering of the control chart dataset. SAX = symbolic aggregate approximation.

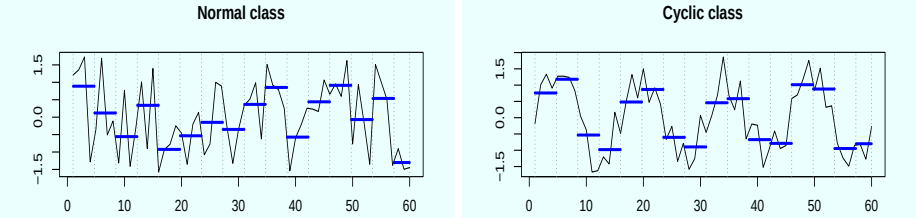

Figure 4.4. Normal and cyclic class converted by the piecewise aggregate approximation.

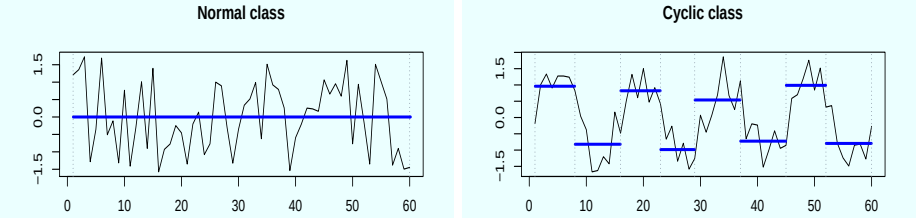

Figure 4.5. Normal and cyclic class converted by the unbalanced Haar wavelet transformation.

5. 결론 및 논의점

본 논문에서는 SAX의 한계점을 보완하기위해 불균형 Haar 웨이블릿 변환을 이용하여 차원을 축소하는 수정된 SAX 방법을 제안하였다. 불균형 Haar 웨이블릿 변환은 자료의 특성에 따라 변화점을 찾기에 차원의 축소 및 보다 정확한 국소 평균 근사값을 제공한다. 불균형 Haar 웨이블릿 변환은 PAA보다 높은 계산 복잡도를 요구하지만 사용자가 임의로 세그먼트의 수(w)을 정하는 SAX의 모호성을 제거하며 자료의 특징을 반영하여 서로 다른 크기의 세그먼트로 나눔으로써 중요한 패턴 정보를 보존한 채 효과적으로 차원을 축소한다. 이는 곧 분류와 군집화의 성능 향상을 가져왔으며 UCR archives으로부터의 자

료 분석을 통해 확인하였다.

제안한 방법은 분류와 군집화의 문제에서 좋은 성능을 보이지만 보완해야 할 사항도 존재한다. 본 논문에서 식 (2.12)의 p값에 대한 구체적인 언급을 하지 않았다. 이는 변화점을 얼마나 민감하게 찾아야 하는 것과 관련이 있다. 본 연구에서는 0.8로 설정한 결과로 시작 및 끝점에서 적어도 20% 이상 떨어진 곳에 변화점이 존재함을 의미한다. 즉 변화점 탐지 논문에서 잘 알려진 바 대로 변화점이 시작 및 끝점에 너무 가까우면 자료의 수가 충분치 않기 때문에 변화점을 찾지 못한다. 따라서 적절한 p값이 선택될 경우 분류와 군집화의 성능은 향상될 수 있으며 이 역시 자료에 의존적으로 해결할 필요가 있다고 본다. 또한 근사된 시계열을 심볼릭 자료로 만들 때 역시 a개의 알파벳 수로 치환이 되는데 이 역시 자료의 성질에 따라 결정한다면 좀 더 좋은 결과를 얻을 수 있을 것이라 본다.

References

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. H. (2015). Time-series clustering - a decade review, *Information Systems*, **53**, 16–38.

Baek, C. and Pipiras, V. (2009). Long range dependence, unbalanced Haar wavelet transformation and changes in local mean level, *International Journal of Wavelets, Multiresolution and Information Processing*, **7**, 23–58.

Chan, K. and Fu, W. (1999). Efficient time series matching by wavelets, *ICDE*, **15**, 126–133.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, **81**, 425–455.

Faloutsos, C., Ranganathan, M., and Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases, *SIGMOD Record*, **23**, 419–429.

Fryzlewicz, P. (2007). Unbalanced Haar technique for nonparametric function estimation, *The Journal of American Statistical Association*, **102**, 1310–1327.

Keogh, E., Chakrabarti, K., Pazzani, M., and Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases, *Journal of Knowledge and Information Systems*, **3**, 263–286.

Lin, J., Keogh, E., Wei, L., and Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series, *DMKD*, **15**, 107–144.

불균형 Haar 웨이블릿 변환을 이용한 군집화를 위한 시계열 표현

이세훈[a] · 백창룡[a,1]

[a]성균관대학교 통계학과

(2018년 8월 8일 접수, 2018년 9월 20일 수정, 2018년 11월 15일 채택)

요 약

시계열 데이터의 분류와 군집화를 효율적으로 수행하기 위해 다양한 시계열 표현 방법들이 제안되었다. 본 연구는 Lin 등 (2007)이 제안한 국소 평균 근사를 이용하여 시계열의 차원을 축소한 후 심볼릭 자료로 이산화하는 symbolic aggregate approximation (SAX) 방법의 개선에 대해서 연구하였다. SAX는 국소 평균 근사를 할 때 등간격으로 임의의 개수의 세그먼트로 나누어 평균을 계산하여 세그먼트의 개수에 그 성능이 크게 좌우된다. 따라서 본 논문은 불균형 Haar 웨이블릿 변환을 통해 국소 평균 수준을 등간격이 아니라 자료의 특성을 반영하여 자료 의존적으로 선택하게 함으로써 시계열의 차원을 효과적으로 축소함과 동시에 정보의 손실을 줄이는 방법에 대해서 제안한다. 제안한 방법은 실증 자료 분석을 통해 SAX 방법을 개선시킴을 확인하였다.

주요용어: 시계열 표현, SAX, 불균형 Haar 웨이블릿 변환, 군집화, 분류

이 논문은 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (NRF-2017R1A1A105000831).

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: crbaek@skku.edu