Introduction to Matrix Profile
행렬 프로파일에 대한 이해

파이널 프로젝트 3주차

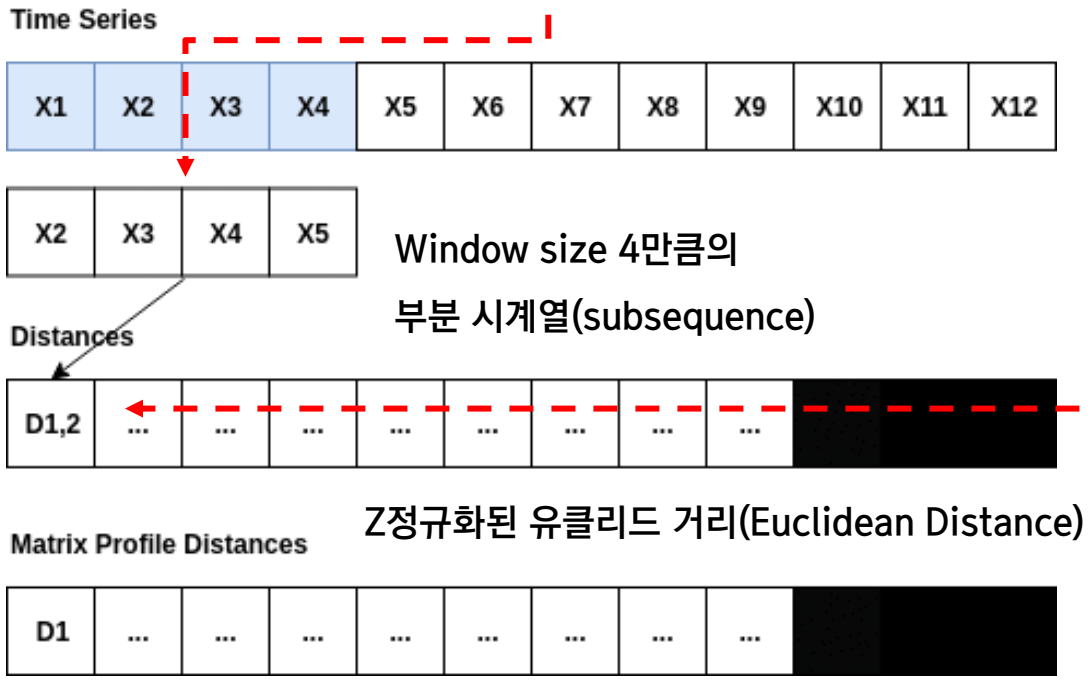
2023.11.22(수) 16:00

김설진/안정현/이승용/이지흔/이혜준

Abstract

Matrix Profile의 아이디어

전체 시계열을 1) 짧은 길이의 부분 시계열로 잘라, 조금씩 이동하여 2) 거리를 측정하는 방법론



Motifs(Patterns)

시계열 내에서 주기적으로 반복되는 패턴, 해당 시계열의 대표성

Discords(Anomalies)

Matrix Profile에서 다른 쿼리들과의 거리값이 비교적 큰 쿼리, 나머지 쿼리들과 가장 상이한 양상을 띄는 것

Matrix Profile

Matrix Profile의 요소

1) 전체 시계열(Time Series) : 시간 T만큼의 시계열 자료

2) 부분 시계열(Subsequence) : i번째 위치에서 m길이 만큼의 값을 가진 sequence

$$T_{i,m} = t_i, t_{i+1}, \cdots, t_{i-m+1},\ 1 \le i \le n-m+1$$

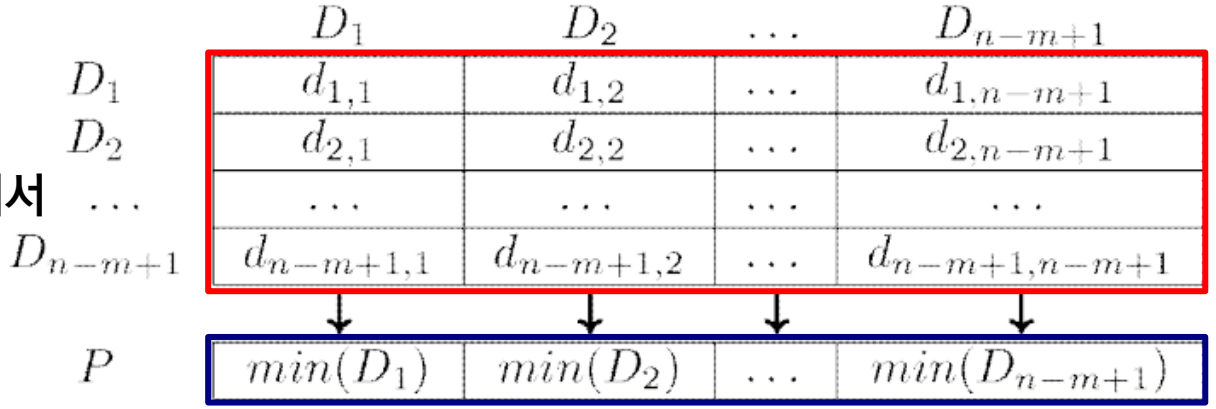
3) 부분 시계열 집합(All-subsequences set) : 전체 시계열 T 중 m을 윈도우 사이즈로 하는 모든 sequence들의 집합

$$A = [T_{1,m}, T_{2,m}, \cdots, T_{n-m+1,m}]$$

4) **거리 행렬(Distance Matrix)** : 모든 sequence 집합에 존재하는 원소들과 기준 subsequence와의 유클리드 거리

- 거리 행렬은 대칭
- 대각행렬은 0 (자기 자신과의 거리는 0)
- 행렬의 크기는 (T-m+1)x(T-m+1)

5) **행렬 프로파일(Matrix Profile)** : 거리 행렬의 각 열에서 가장 작은 값, 즉 자기 자신을 제외한 가장 가까운 subsequence와의 거리를 원소로 가지는 행렬



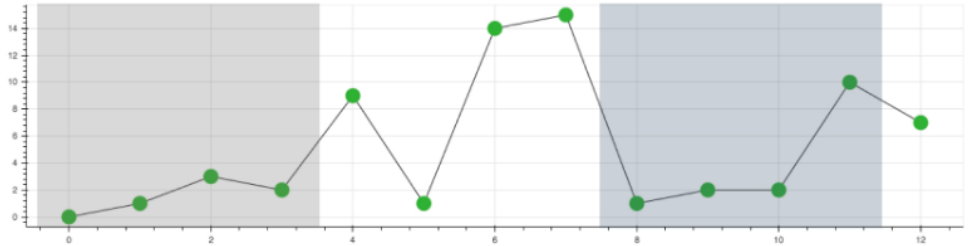
source : 백창룡 and 조예나. (2022). 행렬 프로파일을 이용한 금융 시계열 분석. Journal of The Korean Data Analysis Society, 24(1), 67-81.

Matrix Profile

Matrix Profile의 계산 예시

1) Window size 4만큼의 길이를 가지는 subsequence

0	1	3	2	9	1	14	15	1	2	2	10	7



- 위 그림을 보면 0번째 인덱스에서의 subsequence와 8번째 인덱스에서의 subsequence 간 유사한 패턴이 있음을 확인 가능
- 육안 상으로 비슷해 보일 뿐, 정확한 판단을 위해서는 수치화된 score가 필요

2) 기준 subsequence와의 유클리드 거리 계산

0	1	3	2	9	1	14	15	1	2	2	10	7

- 사이즈 4의 window를 옆으로 sliding하면서 pairwise Euclidean distance를 계산
- 하나의 기준 sequence 당 총 10(13-4+1)개의 거리가 도출되며, 이와 같은 계산을 총 10(13-4+1)번 반복
- (n-m+1)x (n-m+1) 사이즈의 거리 행렬(Distance Matrix) 산출

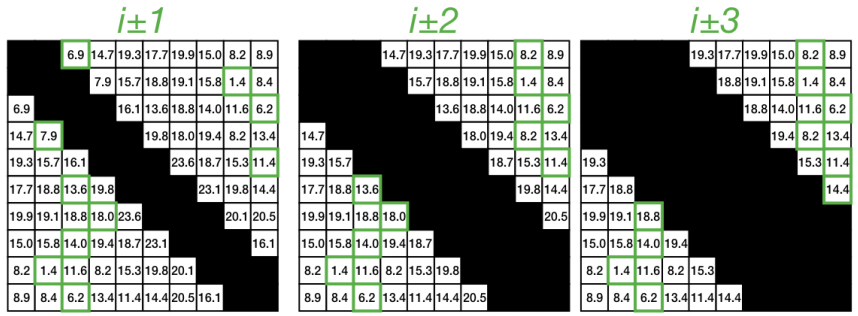
$$\text{Euclidean Distance(d)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

source : Introduction to Matrix Profile

Matrix Profile

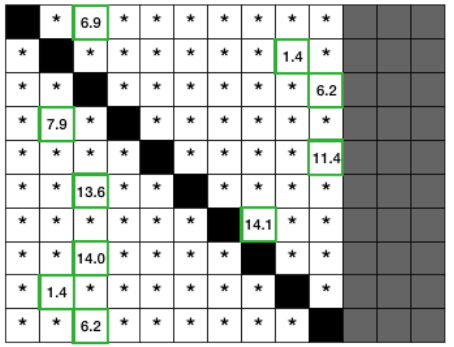
Matrix Profile의 계산 예시

3) Exclusion zone 지정



- exclusion zone : trivial matches(자기 자신 또는 인접한 시계열과의 거리는 필연적으로 매우 가까움)를 피하기 위해 설정
- 일반적으로 현재 window index를 기준으로 window size(m)의 1/2만큼 전후로 하여 exclusion zone 지정
- 위 케이스에서는 자기 자신과의 거리인 0만 제외

4) 거리 행렬에서 행렬 프로파일(Matrix Profile) 추출



- Distance Matrix의 각 행 또는 열에서 최소값만을 저장
- 최소값만을 고려하는 것은 공간 복잡도(메모리 사용량)를 최소화 하는 것을 의미

source : Introduction to Matrix Profile

Matrix Profile

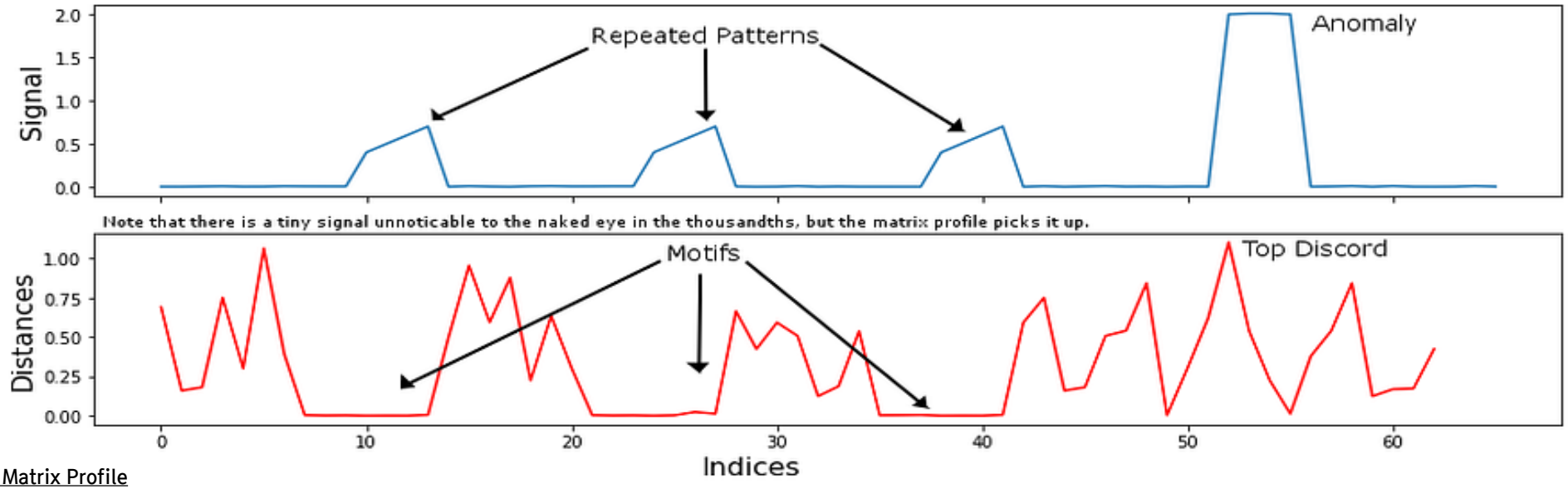
Matrix Profile의 해석

모티프(motifs)

- 시계열 내에서 가장 비슷한 패턴을 가지는 부분 시계열 쌍
- 반복적으로 나타나는 유사한 패턴들의 경우, Matrix Profile에서 낮은 수치를 가짐

디스코드(discords)

- Matrix Profile에서 가장 큰 값이며 이는 유사한 부분 시계열이 없거나 매우 이질적인 패턴이라는 것을 의미
- 이상탐지에 활용 (global discord vs local discord)



source : Introduction to Matrix Profile