Discrete Wavelet Transform을 이용한 Time Series Classification

방태모

Time Series Classification(TSC)는 이미지 또는 음성 인식, 의학적 진단(심전도를 통한 심장병의 유형 진단), gesture detection 등등에서 특히 많이 쓰인다. 하지만 시계열 데이터의 특성이 가진 높은 차원으로 인해, 계산비용이 크다는 문제점을 가지고 있다. 이 문제를 해결하기 위해 효율적인 차원 감소를 위한 방법 중 하나로 Discrete Wavelet Transform(DWT)이 제시된다. DWT가 TSC에서 가지는 이점은 아래와 같다.

① DWT를 통해 시계열에 Feature engineering을 수행함으로써, 시계열 데이터의 차원 축소를 하는 동시에, 원 데이터 셋(the original uncompressed dataset)과 비교하여 비슷하거나 더 좋은 정확도를 얻어낼 수 있다.
② Feature engineering을 통해 불필요한 잡음을 제거할 수 있다.
③ 몇몇 데이터에서는 compression residual details를 사용한 TSC가 원 데이터 또는 DWT를 적용한 데이터보다 훨씬 더 좋은 성능을 보인다. 이는 특정 데이터들은 global features를 사용하는 것 대신 local features를 사용하는 것이 더 좋다는 것을 보여준다.

Wavelet Transform은 Fourier Transform의 발전된 형태라고 할 수 있다. Fourier Transform의 핵심은 아래 식과 같이 아무리 복잡한 형태의 파동이라도 간단한 파동(sin, cos)의 선형결합의 형태로 나타낼 수 있다는 것이다.

any 2π-periodic function $f(x)$ is the sum

$$a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

of its Fourier series. The coefficients a_0, a_k and b_k are calculated by

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(x)dx, \quad a_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(kx)dx, \quad b_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(kx)dx$$

(출처 : < An Introduction to Wavelets - Amara Graps >)

하지만 Fourier Transform의 시간 정보를 고려할 수 없다는 단점이 있으며, 이를 보완하여 만들어진 알고리즘이 Wavelet Transform에 해당한다. 자세한 이해를 위해 기존의 시계열 자료가 갖는 특성, 그리고 Fourier Transform, Short Time Fourier Transform, Discrete Wavelet Transform을 통해 얻어낼 수 있는 특성을 간략하게 그림으로 비교해보자.

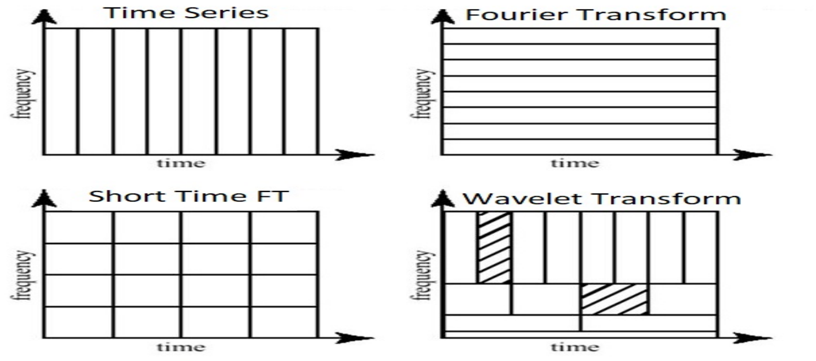


Figure 2. A schematic overview of the time and frequency resolutions of the different transformations in comparison with the original time-series dataset. The size and orientations of the block gives an indication of the resolution size.

(1) Original Time Series
시간 영역 : A high resolution
주파수 영역 : A zero resolution

(2) Fourier Transform(이하, FT)
시간 영역 : A zero resolution
주파수 영역 : A high resolution

(3) Short Time Fourier Transform(이하, STFT)
It has medium sized resolution in both the frequency and time domain.

FT가 시간 정보를 전혀 반영하지 못한다는 것을 극복하기 위한 모티베이션에서 STFT가 나왔으며, 원래의 신호를 FT를 적용하기 전에 같은 길이로 나누는 것이 아이디어이다. 그러나, STFT의 문제는 uncertainty principle이라고 알려진 FT의 이론적 한계에 부딪히게 된다는 것이다. 여기서 말하는 이론적 한계란, FT에서 창을 작게 가져갈수록 해당 신호(시계열)에서 특정 주파수가 어디서 기여하는지는(time) 알 수 있지만, 주파수 값 자체에 대해서는 잘 알지 못하게 되는 문제를 말한다. 반면 창을 크게 가져가면, 주파수 값에 대해서는 잘 파악할 수 있지만 특정 주파수가 어디서 기여하는지에 대해 잘 알 수 없게 된다.

이러한 다이나믹한 주파수 스펙트럼을 가진 신호를 분석하는 데에 더 나은 접근 방법이 바로 Wavelet Transform 이다. Wavelet Transform은 scaling factor를 조절함에 따라서 주파수 영역과 시간 영역 모두에 높은 해상도를 갖게 된다.

(4) Wavelet Transform

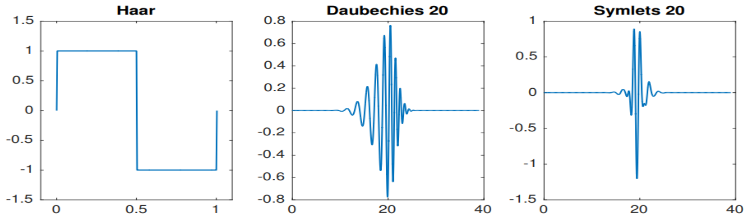
① 작은 주파수 값의 경우
시간 영역 : Low resolution
주파수 영역 : High resolution

② 큰 주파수 값의 경우
시간 영역 : A High resolution
주파수 영역 : A Low resolution

즉, Wavelet Transform은 신호의 특성에 따라 시간 영역, 주파수 영역 각각에 대해 트레이드 오프를 가능하게끔 해준다. 이는 Wavelet Transform의 scaling factor를 통해 조절 할 수 있다. scale은 주파수의 역수로 주기를 의미한다. 예를 들어, Wavelet Transform을 수행할 때 scaling factor 값을 키우면(longer wavelet) 더 작은 주파수를 분석하게 되어 주파수

영역에 high resolution을 가지게 된다. 반대로, scaling factor 값을 작게 가져가면, 시간 영역에서 더 상세한 정보를 갖게 된다.

아울러, Fourier Transform은 basis function으로 sin, cos 함수만을 사용하지만, Wavelet transform은 아래와 같은 다양한 basis function을 사용할 수 있다.



출처 : < Time Series Classification with Discrete Wavelet Transformed Data : Insights from an Emprical Study - Daoyuan Li >)

위 세 함수 외에도 많은 wavelet function이 존재하며, 세가지만 소개한다. Haar wavelet은 매우 간단하다는 장점이 있지만, 함수가 많이 매끄럽지 않아 범용성이 낮기 때문에(has a low regularity) 교육목적으로 많이 쓰인다. Daubechies 20과 Symlets 20은 Haar보다 더 복잡하여 신호를 좀더 정확하게 표현할 수 있기때문에 범용성이 더 높다(have a higher regularity). Haar transform은 주어진 시계열을 approximations, details로 두 시계열로 분해한다. approximations는 전체적인 shape(global features)를 잡아내고, details는 시계열의 변동(local features)을 나타낸다. 아래 그림은 Haar transform을 적용한 예시이다.

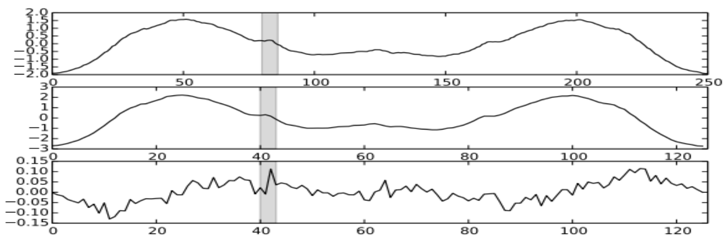


Figure 3: Example of Haar transform: the original signal, the Haar approximation and the residual details.

출처 : < Time Series Classification with Discrete Wavelet Transformed Data : Insights from an Emprical Study - Daoyuan Li >)

Haar transform이 적용된 계열은 원 계열의 size의 절반만을 가지고 있고, 진폭은 조금 커졌지만 원 계열의 shape를 그대로 가지고 있다.