J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

이상탐지 기반의 효율적인 시계열 유사도 측정 및 순위화☆

Efficient Time-Series Similarity Measurement and Ranking Based on Anomaly Detection

최 지 현[1] 안 현[1*]
Ji-Hyun Choi Hyun Ahn

요 약

시계열 분석은 시간 순서로 정렬된 데이터로부터 다양한 정보와 인사이트를 발견하기 위한 방법으로 많은 조직에서 비즈니스 문제 해결을 위해 적용하고 있다. 그중에서 시계열 유사도 측정은 패턴이 비슷한 시계열들을 식별하기 위한 단계로서 시계열 검색 및 군집화와 같은 시계열 분석 응용에서 매우 중요하다. 본 연구에서는 전체 시계열이 아닌 이상치들을 중심으로 시계열 유사도 측정을 계산 효율적으로 수행하는 방법을 제안한다. 이와 관련하여 이상탐지를 통해 추출된 서브시퀀스 집합에 대한 유사도 측정 결과와 시계열 전체에 대한 유사도 측정 결과 사이의 순위 상관관계를 측정 및 분석하여 제안 방법을 검증한다. 실험 결과로써, 주식 종목 시계열 데이터에 이상치 비율 10% 을 적용한 유사도 측정으로부터 최대 0.9 이상의 스피어만 순위 상관계수를 확인하였다. 결론적으로 제안 방법을 통해 시계열 유사도 측정에 소요되는 계산량을 유의미하게 절감하는 동시에 신뢰 가능한 시계열 검색 및 군집화 결과를 기대할 수 있다.

☞ 주제어 : 시계열 유사도, 이상탐지, 서브시퀀스, 스피어만 순위 상관계수

ABSTRACT

Time series analysis is widely employed by many organizations to solve business problems, as it extracts various information and insights from chronologically ordered data. Among its applications, measuring time series similarity is a step to identify time series with similar patterns, which is very important in time series analysis applications such as time series search and clustering. In this study, we propose an efficient method for measuring time series similarity that focuses on anomalies rather than the entire series. In this regard, we validate the proposed method by measuring and analyzing the rank correlation between the similarity measure for the set of subsets extracted by anomaly detection and the similarity measure for the whole time series. Experimental results, especially with stock time series data and an anomaly proportion of 10%, demonstrate a Spearman's rank correlation coefficient of up to 0.9. In conclusion, the proposed method can significantly reduce computation cost of measuring time series similarity, while providing reliable time series search and clustering results.

☞ keyword : Time-series Similarity, Anomaly Detection, Subsequences, Spearman's Rank Correlation Coefficient

1. 서 론

시계열은 시간에 따라 순차적으로 측정한 값의 집합으로서, 시계열 패턴 분석 및 예측 기법에 기반한 다양한 응용을 가진다. 이와 관련하여 시계열 유사도 측정은 시계열 군집화 및 검색에 필수적인 단계로서 시계열의 관측치, 모양, 경향 등에 기반하여 시계열 사이의 유사성을 정량화한다.

일반적으로 시계열 유사도는 두 시계열의 관측치 차이에 거리 함수를 작용하여 측정한다. 그러나, 최근에는 사물인터넷 및 빅데이터 기술의 발전으로 다양한 분야에서 데이터들이 방대하게 생성되고 있으므로, 시계열 유사도 측정 또한 계산 비용 측면에서 효율성을 고려해야 한다. 이를 위해, 압축[4, 5] 또는 전체 시계열이 아닌 부분 시계열(이하 서브시퀀스[6, 9]) 기반의 시계열 유사도 측정 방법들이 제안되었다.

본 논문에서는 이상탐지 기법에 의해 추출된 서브시퀀스 집합을 대상으로 시계열 유사도를 효율적으로 측정하는 방법을 제안한다. 이상치는 일반적인 데이터 패턴

1 Department of Information and Communication, Hanshin University, Gyeonggi, 18101, Korea.
* Corresponding author (hyunahn@hs.ac.kr)
[Received 29 January 2024, Reviewed 14 February 2024, Accepted 19 February 2024]
☆ 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. NRF-2021R1F1A1063160)과 한신대학교 학술연구비 지원에 의하여 수행되었음

에서 벗어난 값으로서 제거의 대상으로 고려되지만, 본 연구에서는 시계열들을 구별하기 위한 효과적인 특징으로서 활용한다. 즉, 이상탐지에 의해 식별된 서브시퀀스 집합에 해당되는 데이터 패턴이 시계열별로 유의미한 차이를 가진다는 것이 본 연구의 주요한 가정이다. 제안 방법은 선택된 기준 시계열의 이상치들에 해당되는 서브시퀀스 집합을 추출하여, 이들을 비교 대상 시계열들과 일대다(one-to-many) 방식으로 유사도를 측정한다.

논문의 구성은 다음과 같다. 2장에서는 시계열 유사도 관련 연구들을 요약하며, 3장에서는 본 연구에서 활용하는 이상탐지 기법에 대해 설명한다. 4장에서는 제안 방법 및 관련 성능 지표에 대해 설명하며, 마지막 5장에서는 본 연구의 한계점 및 향후 연구 계획으로 결론을 맺는다.

2. 관련 연구

시계열은 기본적으로 추세(trend), 계절성(seasonality), 잔차(residual) 성분으로 구성되며[1], 시계열 사이의 상관성 또는 시차 등의 다양한 특성을 가지므로 이를 고려하여 시계열 유사도를 측정하는 것이 중요하다.

유클리드 거리는 시계열 유사도 측정을 위한 기본적인 기준으로서 같은 시점에서의 두 시계열 관측치 사이의 거리를 의미한다. 직관적이지만 스케일 차이가 크거나 이상치에 취약하다는 단점이 있다. 또한 길이가 서로 다른 시계열에는 적용할 수 없으며, 시계열 사이의 시차를 고려하지 않는다는 한계점이 존재한다. 이를 보완하기 위해 제안된 ERP(Edit distance with Real Penalty[2])는 특정 시점에서 두 시계열 모두 관측치를 포함하는 경우 유클리드 거리를 적용하며, 그렇지 않은 경우 일정의 페널티를 부여함으로써 길이가 상이한 시계열에 대해서도 유사도 측정이 가능한 방법이다. 반면에, DTW(Dynamic Time Warping[3])는 특정 시점과 주변 시점의 거리를 기반으로 거리 누적합이 최소화되는 방향으로 두 시계열의 각 시점을 매핑하기 때문에 시차를 반영한 유사도 측정이 가능하다. 대응되는 모든 시점 쌍에 대한 연산을 수행하기 때문에 높은 정확도를 기대할 수 있지만 많은 계산량을 요구하는 문제점이 있다.

이와 관련하여 시계열 유사도를 효율적으로 측정하기 위해 압축 기반의 유사도 측정 방법들이 제안되었다. PAA(Piecewise Aggregate Approximation[4])은 시계열을 일정 구간으로 나눠 구간의 평균값을 대표값으로 적용하여 시계열을 단순화한다. 이와 유사하게, SAX(Symbolic Aggregate Approximation[5])는 PAA 방법을 기반으로 실수로 구성된 각 구간의 대표값을 범주형 값으로 변환하여 탐색 공간을 효과적으로 줄인다.

한편, 전체 데이터가 아닌 부분 집합인 서브시퀀스 단위로 유사도를 효율적으로 측정하기 위한 방법들이 제안되었다. LCSS(Longest Common Subsequence[6])는 각 시점 별로 두 시퀀스의 관측치 거리가 기준 임계값 보다 작으면 1, 크면 0으로 표현하여 연속적인 1이 최대가 되는 서브시퀀스를 찾는 알고리즘으로, 시계열 유사도 측정 문제에서 빈번히 활용된다[7, 8]. 이 방법은 이상치에 강건하고, 결측치를 포함하거나 길이가 서로 다른 시계열에도 적용 가능한 장점을 가지지만 수동적으로 적합한 임계값을 설정하는 어려움이 있다.

MVM(Minimal Variance Matching[9])은 두 시계열이 대응하는 시점에서의 거리의 합을 최소화 하기 위해 차분 행렬을 구성하고, 이를 통해 최소 비용의 DAG(Directed Acyclic Graph)를 찾아 유사도를 측정한다. 값의 차이를 기반으로 작동하기 때문에 임계값이 요구되지 않지만 이상치에 민감하다.

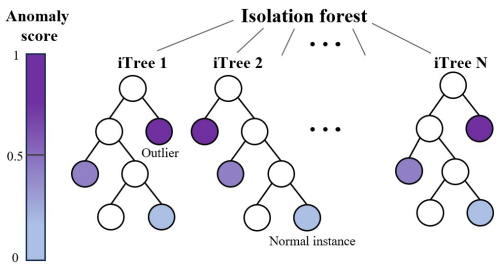
위의 연구들과 다르게, 본 연구에서는 계산 대상을 이상탐지에 의해 식별된 서브시퀀스 집합으로 제한하여 시계열 유사도를 효율적으로 측정하는 방법을 제안한다.

3. 이상탐지 기법

본 장에서는 시계열 유사도 측정 대상이 되는 서브시퀀스 집합을 식별하기 위한 두 가지 이상탐지 기법에 대해 설명한다.

3.1 Isolation Forest

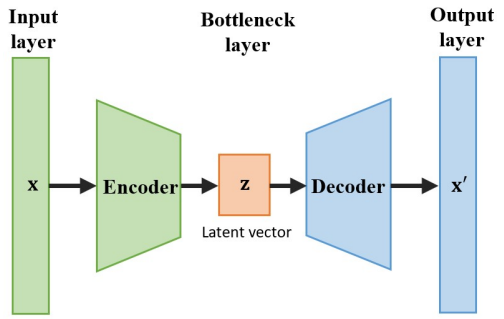
Isolation Forest[10]는 밀도 기반 이상탐지 기법으로서, 무작위로 선택된 변수와 분할 값을 기준으로 데이터를 반복적으로 분할하는 방식으로 다수의 이진 탐색 트리(binary search tree)를 생성한다. 이 방법의 주요 가정은 이상치들은 데이터 분할 과정에서 쉽게 고립되며, 정상 데이터들은 잘 고립되지 않는 것이다. 각 트리의 분할 과정은 모든 데이터들이 개별적으로 고립될 때까지 수행되며, 개별 데이터의 이상 점수(anomaly score)는 해당 데이터에 대응되는 노드의 경로길이 평균(또는 중앙값)을 기반으로 계산된다. 상대적으로 계산 비용이 낮기 때문에 다변량 데이터에 대한 이상탐지에 많이 적용된다.



(그림 1) Isolation Forest의 개념
(Figure 1) The concept of Isolation Forest

3.2 Autoencoder

Autoencoder[11]는 데이터를 압축하고 복원하는 방식으로 동작하는 딥러닝 계열 모델로서 크게 인코더(encoder)와 디코더(decoder) 두 부분으로 구성된다. 인코더는 입력 데이터를 저차원으로 압축하는 역할을 수행하며, 이 과정에서 원본 데이터의 특징을 처리하여 저차원의 잠재 벡터(latent vector)로 변환한다. 디코더는 잠재 벡터를 다시 고차원 데이터로 복원하여 재구성한다. 기본적으로 입력 계층과 출력 계층의 차원이 같고, 중간 계층(bottleneck layer)의 차원이 적은 심층 신경망 구조를 가진다.



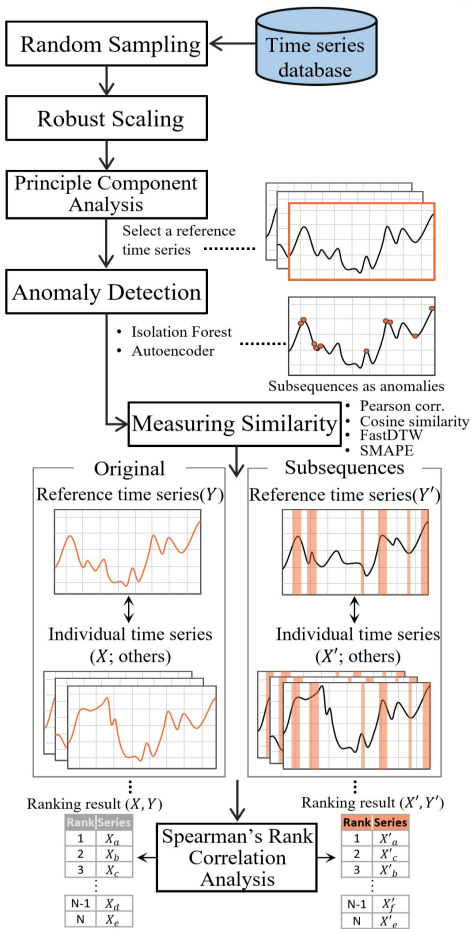
(그림 2) Autoencoder의 구조
(Figure 2) The architecture of Autoencoder

Autoencoder의 주된 목표는 재구성 오차(reconstruction error)를 최소화하면서 동시에 잠재 표현이 입력 데이터의 중요 특징을 포함하도록 하는 것이며, 이를 통해 데이터의 차원을 축소하거나 중요한 특징을 추출하는 데에 활용할 수 있다. 특히, 이상치의 경우 Autoencoder의 압축-복원 과정에 의해 처리된 결과가 일반적으로 높은 재구성 오차를 보일 것으로 예상되므로 본 연구에서는 Autoencoder 모델을 이상탐지 기법으로써 활용한다.

4. 이상탐지 기반 시계열 유사도 측정

4.1 제안 방법

본 연구에서 제안하는 이상탐지 기반 시계열 유사도 측정의 전체 수행 과정은 그림 3과 같다.



(그림 3) 제안 방법의 수행 과정
(Figure 3) The overall procedure of the proposed method

먼저, 무작위로 선택된 다중 시계열 데이터를 Robust Scaling 방식으로 정규화하고, 주성분 분석(PCA: Principle Component Analysis[12])을 통해 차원 축소를 수행한다. 다음으로 특정 기준 시계열(reference time series, Y)을 선택하고 이에 대해 이상탐지를 수행하여 이상치에 해당되

는 시간 인덱스 집합을 추출한다. 이를 기준 시계열 및 비교 대상 시계열에 적용하여 서브시퀀스 집합(Y', X')을 각각 추출한 뒤 시계열 유사도를 측정한다. 마지막으로, 원본 시계열(Y, X)과 서브시퀀스 집합(Y', X')의 유사도 측정 결과에 스피어만 순위 상관계수(Spearman's Rank Correlation Coefficient)를 적용하여 제안 방법을 검증한다.

4.2 차원 축소

고차원 데이터 분석에는 모델링의 복잡성 및 계산 비용의 증가와 같은 문제가 있으며, 이상치가 존재할 경우 전체적인 데이터의 경향성을 설명하지 못하고 왜곡시킬 수 있다. 그러므로, 본 연구에서는 다변량 시계열 데이터의 차원 축소를 위해 주성분 분석을 적용한다. 이는 데이터의 분산을 최대한으로 보존하는 주성분들을 찾으며 일반적으로 누적기여율(cumulative proportion)이 85% 이상을 만족할 때의 주성분들로 선택하여 차원 축소를 수행한다.

4.3 시계열 유사도

본 절에서는 제안 방법에서 적용한 네 가지 시계열 유사도 측정 방법에 대해 설명한다.

4.3.1 코사인 유사도

코사인 유사도(cosine similarity)는 내적 공간에서 두 변수의 사이각을 이용해 유사도를 측정하는 지표이다. 정규화된 두 변수의 절대적인 사이각을 기준으로 유사도를 측정하기 때문에 변수 간의 크기 차이에 영향을 받지 않으며 방향성을 고려하여 유사도를 측정한다.

$$\cos(x,y) = \frac{xy}{\| x \| \, \| y \|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

수식 1과 같이 코사인 유사도는 두 벡터 x, y의 내적을 두 변수의 크기의 곱으로 나눈 값이다. 측정된 유사도는 [-1, 1]의 범위를 가지며 1에 가까울수록 두 변수는 높은 유사성을 가진 것으로 해석한다. 본 연구에서는 n의 길이를 갖는 기준 시계열 및 비교 대상 시계열 내의 두 변수를 벡터 x와 y로 대입하여 코사인 유사도를 측정한다.

4.3.2 피어슨 상관계수

피어슨 상관계수(Pearson correlation coefficient)는 기준 벡터의 값이 증가 또는 감소할 때 다른 벡터의 변화를 기준으로 두 벡터의 선형 상관관계를 측정하는 지표이다. 각 벡터에 표준화된 값을 사용하기 때문에 서로 다른 범위의 값을 가진 벡터에 대해 영향을 받지 않고 유사도를 측정할 수 있다.

$$r(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (2)$$

n은 시계열의 길이를, $\bar{x}$와 $\bar{y}$는 각 벡터의 평균값을 의미한다. 상관계수가 1에 가까울수록 두 벡터는 강한 양의 상관관계를 가지며 -1에 가까울수록 강한 음의 상관관계를 가지는 것으로 해석한다. 반면, 상관계수가 0에 가까울수록 두 벡터 사이에 유의미한 상관관계가 없다는 것을 의미한다.

4.3.3 FastDTW

FastDTW(Fast Dynamic Time Warping[13])는 두 시계열의 시차 또는 왜곡된 패턴들을 고려하여 유사도를 측정하는 DTW 알고리즘의 계산 복잡성을 개선한 모델이다. PAA 알고리즘을 통해 고차원 시계열을 저차원으로 매핑하여 DTW 거리 행렬을 구성하여 최적의 경로를 탐색한다. 이를 통해 데이터의 핵심 특성을 유지하면서 전체 계산 비용을 감소시켜 효율적으로 유사도를 측정한다.

$$D(i,j) = dist(x_i, y_j) + \min \begin{pmatrix} D(i-1,j) \\ D(i,j-1) \\ D(i-1,j-1) \end{pmatrix} \quad (3)$$

$D(i,j)$는 거리 행렬에서의 누적 거리를 구하는 수식으로, DTW와 동일한 계산 방식을 사용한다. 간단히 설명하면, 거리 함수 $dist(x_i, y_j)$를 사용하여 두 시계열의 거리를 계산하고, 이전 타임 스텝에서 구해진 누적 거리 3개 중 최소값을 선택해 더하는 과정을 반복하여 거리 행렬을 구성한다. 최종적으로 구성된 거리 행렬에서 누적 거리의 합이 최소가 되는 최적의 경로를 찾는다.

4.3.4 SMAPE

SMAPE(Symmetric Mean Absolute Percentage Error)는 관측값과 예측값 사이의 상대적 오차에 기반하여 회귀모형의 성능을 측정하는 지표이다. 절대 오차를 절대 관측값과 절대 예측값의 평균으로 나누기 때문에 관측값이 0인 경우에도 zero division 문제가 발생하지 않는다. SMAPE는 [0, 200]의 범위를 가지며, 값이 작을수록 모델의 예측이 정확한 것으로 해석된다. 본 논문에서는 zero division 문제를 방지하고 백분율 형태의 시계열 유사도를 측정하기 위해 SMAPE를 적용한다.

$$SMAPE(x,y) = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - x_i|}{(|y_i|+|x_i|)/2} \quad (4)$$

즉, 시점 i에서의 기준 시계열의 벡터 y의 관측값을 y_i로, 비교 대상 시계열의 벡터 x의 관측값을 x_i로 대입하여 유사도를 측정한다.

4.4 성능 지표

제안 방법의 정확성을 평가하기 위해 본 논문에서는 순위 척도로 표현된 변수 간 상관관계를 분석하기 위한 스피어만 순위 상관계수를 적용한다. 원본 데이터셋에서 i번째 시계열에 대한 유사도 순위를 $R(X_i)$로, 이상탐지에 의해 식별된 서브시퀀스 집합에 대한 유사도 순위를 $R(X_i')$로 정의한다. 이를 기반으로 스피어만 상관계수는 다음 수식에 따라 측정된다.

$$\rho = 1 - \frac{6\sum d_i^2}{N(N^2-1)} \quad (5)$$

N은 시계열의 개수이며, d_i는 원본 시계열과 서브시퀀스 집합에서의 i번째 시계열의 유사도 순위 차이를 나타내며 $R(X_i) - R(X_i')$로 정의된다. 측정된 상관계수 ρ가 1에 가까울수록 두 순위 결과 사이의 강한 양의 상관관계가 있는 것으로 해석하므로, 제안 방법에 의한 유사도 측정 결과가 정확한 것으로 판단할 수 있다.

5. 성능 평가

본 장에서는 NASDAQ 종목의 주가 데이터를 이용한 시계열 유사도 측정에 대한 실험 결과를 기술한다.

5.1 실험 방법

본 연구에서는 두 가지 이상탐지 기법(Isolation Forest, Autoencoder) 및 이상치 비율(0.05, 0.1, 0.15, 0.25), 그리고 PCA 적용 여부를 조합하여 실험을 수행하였다. 또한, 조합별 실험에서 네 가지 유사도(코사인 유사도, 피어슨 상관계수, FastDTW, SMAPE)를 측정하여 순위화한 결과에 스피어만 순위 상관계수를 각각 측정하여 비교하였다.

5.2 데이터셋

실험을 위한 예제 데이터로써 금융 데이터 라이브러리인 Finance Data Reader[14]와 NASDAQ 공식 웹사이트를 통해 수집한 주가 데이터를 사용하였다. 전체 종목 중에서 충분한 데이터 확보와 시장 영향력을 고려하여 5년 이상의 데이터가 확보된 주식 종목 3,915개를 선별하였다. 데이터의 샘플링 단위는 1일이며, 총 6개의 변수를 가진다. 각 변수에 대한 설명은 아래와 같다.

(표 1) 데이터셋 변수 요약
(Table 1) Summary of the variables in the dataset

변수	설명	자료형
Open	시가(Opening price)	float
High	고가(High price)	float
Low	저가(Low price)	float
Close	종가(Closing price)	float
Adj Close	수정 종가(Adjusted closing price)	float
Volume	해당 개장일의 거래량	int

실험에서는 기준 시계열로 활용되는 섹터별 시가총액 상위 11개 종목을 포함한 총 100개의 주가 데이터를 선택하여 유사도 측정을 수행하였다.

(표 2) 섹터별 시가총액 기준 상위 종목(11개)
(Table 2) Top 11 stocks by market capitalization by sector

순번	심볼	섹터	산업 부문
1	AAPL	기술	컴퓨터 제조업
2	CSCO	통신	컴퓨터 통신 장비
3	AMGN	헬스케어	생물학적 약제
4	CME	금융	투자은행/브로커/서비스
5	EQIX	부동산	부동산 투자 신탁
6	AMZN	자유 소비재	카탈로그/전문 유통
7	PEP	필수 소비재	음료(생산/유통)
8	HON	산업	항공 우주
9	UFPI	원자재	임산공업
10	FANG	에너지	기름·가스 생산
11	EXC	유틸리티	발전

5.3 데이터 전처리

종목별 상이한 주식 가격의 범위를 정규화하기 위해 Robust scaling을 사용하였다. Robust scaling은 사분위수 범위(IQR: Inter Quartile Range)와 중앙값(median)을 사용해 선형적으로 값을 변환하는 정규화 방법이다. 최소-최대 정규화(min-max normalization) 또는 분산을 사용하는 표준화 방식의 정규화와는 다르게 중앙값을 사용하기 때문에 이상치의 영향력을 최소화한다. 벡터 x에 대한 Robust scaling은 아래와 같이 계산된다.

$$x_{scaled} = \frac{x - median(x)}{IQR(x)} \quad (6)$$

x_{scaled}는 관측 값 x의 정규화 결과로서 x에 중앙값을 뺀 값을 해당 데이터의 Q3(하위 75%)과 Q1(하위 25%)의 차이에 해당되는 사분위수범위로 나눈 값이다.

정규화를 통해 변환된 데이터의 차원 축소를 위해 그림 3과 같이 PCA를 적용하였다. 표 3은 헬스케어 섹터의 대표 종목인 Amgen의 데이터에 주성분 분석을 적용한 결과를 나타낸다. 주성분 기여율은 0.6538, 0.3383, 0.0064, 0.0008, 0.0004, 0.0001이며, 일반적으로 누적 기여율 85% 이상일 때의 주성분을 선택하는 기준에 따라 두 번째 주성분까지 적용해서 차원 축소된 데이터를 얻었다(누적 기여율: 0.9921). 제안 방법의 마지막 단계인 스피어만 순위 상관계수 측정이 변수 단위의 연산이므로, 주성분 별로 수행된다. 본 연구에서는 주성분 기여율을 반영하여 순위 상관계수를 측정하기 위해 다음의 수식을 적용한다.

$$\rho^S = \sum_{i=1}^{2} \rho_i w_i \quad (7)$$

ρ^S는 스피어만 순위 상관계수 측정의 최종 결과로서 개별 변수의 스피어만 상관 계수 ρ_i와 주성분 기여율 w_i를 곱한 값의 총합이다.

(표 3) 특정 종목(심볼: AMGN)의 주성분 분석 결과
(Table 3) Results of principal component analysis of the specific stock(symbol: AMGN)

주성분 순번	고유값	기여율	누적 기여율
1	2.6766	0.6538	0.6538
2	1.3851	0.3383	0.9921
3	0.0263	0.0064	0.9985
4	0.0035	0.0008	0.9994
5	0.0017	0.0004	0.9998
6	0.0005	0.0001	1.0000

5.4 실험 결과

표 4는 본 연구의 실험 결과로서 각 이상탐지 방법 별

(표 4) 스피어만 순위 상관계수 측정 결과
(Table 4) Measurement results of the Spearman's rank correlation coefficient

이상치 비율	이상탐지 방법	Cosine	Pearson	FastDTW	SMAPE
0.05	Isolationforest + PCA	**0.8725**	0.8121	0.6532	0.8011
	Autoencoder + PCA	0.8429	0.7213	0.6807	0.8387
	Isolationforest	0.8504	0.8435	0.8238	0.8803
	Autoencoder	0.7191	0.6667	0.6367	0.7478
0.1	Isolationforest + PCA	**0.9143**	0.8984	0.6376	0.8801
	Autoencoder + PCA	0.8922	0.8585	0.7098	0.8635
	Isolationforest	0.8876	0.8821	0.8609	0.8988
	Autoencoder	0.7631	0.7663	0.7569	0.8186
0.15	Isolationforest + PCA	**0.9415**	0.9304	0.7824	0.9109
	Autoencoder + PCA	0.8995	0.8825	0.7046	0.916
	Isolationforest	0.9075	0.9022	0.8779	0.9164
	Autoencoder	0.7862	0.7742	0.7926	0.8184
0.25	Isolationforest + PCA	**0.9615**	0.9524	0.7846	0.9408
	Autoencoder + PCA	0.9144	0.9011	0.6974	0.9356
	Isolationforest	0.9342	0.9293	0.9109	0.9387
	Autoencoder	0.8295	0.7973	0.8205	0.8361

로 측정한 스피어만 순위 상관 계수(ρ^S)를 나타낸다. 측정된 ρ^S가 1에 가까울수록 이상탐지 기반으로 추출된 서브시퀀스 집합을 대상으로 유사도를 측정 및 순위화한 결과와 원본 시계열들을 대상으로 유사도를 측정 및 순위화한 결과 사이에 일치성이 높다는 것을 의미한다.

먼저, 이상탐지 기법 측면에서 Isolation Forest가 Autoencoder와 비교하여 대부분의 경우에서 더 높은 상관 계수를 보였다. 이는 Autoencoder가 고차원 데이터 모델링에 적합한 딥러닝 기법인 반면에, 실험에서 사용한 주가 데이터의 경우 6차원이므로 모델 학습에 충분하지 않았던 것으로 판단된다. 또한, PCA를 적용하여 차원 축소한 데이터를 사용했을 때 더 높은 상관계수가 측정되는 경향을 보였다.

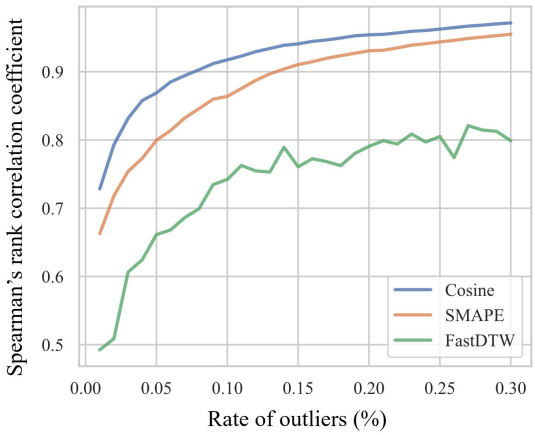
유사도 측정 방법 별로 계산한 ρ^S의 평균은 SMAPE(0.8714), 코사인 유사도(0.8698), 피어슨 상관계수(0.8449), FastDTW(0.7582)의 순서인 반면에, 이상치 비율에 따라 구분된 실험 그룹에서 가장 높은 상관계수는 모두 코사인 유사도를 적용한 경우에서 측정되었다.

전체 실험 결과 중에서는 이상치 비율 0.25에서 Isolation Forest와 PCA, 그리고 코사인 유사도를 적용했을 때 최대값인 0.9615의 상관계수가 측정되었다. 표 5의 스피어만 순위 상관계수의 평가 기준에 따라, 제안 방법에 의한 순위화 결과가 원본 시계열에 대한 유사도 측정 결과에 의한 순위화 결과와 매우 강한 상관관계를 가지는 것을 알 수 있다.

(표 5) 스피어만 순위 상관계수의 평가 기준
(Table 5) Evaluation criteria for Spearman's rank correlation coefficient

상관계수 범위	결과 해석
±0.9 이상	매우 강한 상관관계
±0.7 이상 ±0.9 미만	강한 상관관계
±0.5 이상 ±0.7 미만	중간 정도의 상관관계
±0.3 이상 ±0.5 미만	약한 상관관계
±0.3 미만	매우 약한 상관관계

종합적으로 가장 높은 상관계수를 보이는 방법(Isolation Forest + PCA)에 이상치 비율을 [0, 0.3] 구간으로 세부적으로 나눠서 상관계수를 측정한 결과는 그림 4와 같다. 코사인 유사도 기준으로 이상치 비율이 10%에 근접하면서부터 0.9 이상의 상관계수가 측정되며, 이는 시계열의 전체 데이터가 아닌 서브시퀀스 집합에 대해서 측정하여도 높은 정확도의 유사도를 계산 효율적으로 측정할 수 있음을 나타낸다.



(그림 4) 이상치 비율에 따라 측정된 스피어만 순위 상관계수
(Figure 4) Measured Spearman's rank correlation coefficients by outlier rate

위의 결과에도 불구하고, 본 연구는 몇 가지 한계점을 가진다. 첫 번째, 다변량 데이터를 활용하였지만 변수의 수가 6개로 적기 때문에, 딥러닝 계열의 모델인 Autoencoder에는 적합하지 않았다. 향후에는 추가적인 이상탐지 모델과 고차원의 다중 시계열 데이터셋을 고려하여 제안 방법을 검증하고자 한다.

6. 결 론

본 논문에서는 이상탐지를 통해 식별된 서브시퀀스 집합을 대상으로 시계열 유사도를 효율적으로 측정하는 방법을 제안하였다. 먼저, PCA를 적용하여 차원 축소된 데이터에 대해 기준 시계열을 선정하고 Isolation Forest와 Autoencoder를 통해 이상탐지를 수행하였다. 식별된 이상치에 해당되는 시간 인덱스를 적용하여 기준 시계열과 비교 대상 시계열로부터 서브시퀀스 집합을 각각 추출하여 유사도를 측정하였다. 성능 평가를 위해 유사도 측정 및 순위화의 결과를 제안 방법과 전체 데이터 대상의 방법에서 각각 도출하여 두 결과 사이의 스피어만 순위 상관계수를 측정하였다. 실험 결과, 전체 데이터 대비 적은 비율의 서브시퀀스 집합만으로도 높은 스피어만 순위 상관계수를 도출하였다. 예를 들면, Isolation Forest 이상탐지 기법에 0.1의 이상치 비율을 적용하여 탐지한 서브시퀀스 집합에 코사인 유사도를 적용했을 때, 0.9143의 높은 순위 상관계수를 보였으며, 이는 원본 데이터의 0.1%만을 처리했음에도 제안 방법이 유의미한 유사도 측정

결과를 보였음을 나타낸다. 결론적으로 제안 방법은 주가 분석[15] 및 라이프로그 분석[16]과 같은 영역에서 유사 시계열 검색 및 군집화와 같은 작업에 효율적으로 적용 가능할 것으로 기대된다.

본 연구에서는 주가 데이터에 한정되어 실험을 수행했기 때문에 향후에는 M4[17] 등의 대규모 다중 시계열 데이터셋을 포함한 성능 평가를 추가적으로 수행하여 제안 방법을 더 심도 있게 검증하고자 한다.

참고문헌(Reference)

[1] Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I., "STL: A seasonal-trend decomposition," Journal of Official Statistics, Vol. 6, No. 1, pp. 3-73, 1990.
https://www.math.unm.edu/~lil/Stat581/STL

[2] Chen, L. and Ng, R., "On the marriage of lp-norms and edit distance," Proceedings of the Thirtieth International Conference on Very Large Data Bases, Vol. 30, pp. 792-803, 2004.
https://dl.acm.org/doi/10.5555/1316689.1316758

[3] Muller, M., Information Retrieval for Music and Motion, pp. 69-84, Springer, 2007.
https://doi.org/10.1007/978-3-540-74048-3_4

[4] Keogh, E., Chakrabarti, K., Pazzani, M., and Mehrotra, S., "Dimensionality reduction for fast similarity search in large time series databases," Knowledge and Information Systems, Vol. 3, pp. 263-286, 2001.
https://doi.org/10.1007/PL00011669

[5] Lin, J., Keogh, E., Lonardi, S., and Chiu, B., "A symbolic representation of time series, with implications for streaming algorithms," Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 2-11, 2003. https://doi.org/10.1145/882082.882086

[6] Hirschberg, D. S., "A linear space algorithm for computing maximal common subsequences," Communications of the ACM, Vol. 18, No. 6, pp. 341-343, 1975.
https://doi.org/10.1145/360825.360861

[7] Vlachos, M., Kollios, G. and Gunopulos, D., "Discovering similar multidimensional trajectories," Proceedings 18th International Conference on Data Engineering, pp. 673-684, 2002.
https://doi.org/10.1109/ICDE.2002.994784

[8] Khan, R., et al., "LCSS-based algorithm for computing multivariate data set similarity: A case study of real-time WSN data," Sensors, Vol. 19, No. 1, p. 166, 2019.
https://doi.org/10.3390/s19010166

[9] Latecki, L. J., et al., "Elastic partial matching of time series," Knowledge Discovery in Databases: PKDD 2005: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 577-584, 2005.
https://doi.org/10.1007/11564126_60

[10] Liu, F. T., Ting, K. M., and Zhou, Z. H., "Isolation forest," 2008 Eighth IEEE International Conference on Data Mining, pp. 413-422, 2008.
https://doi.org/10.1109/ICDM.2008.17

[11] Yin, C., Zhang, S., Wang, J., and Xiong, N. N., "Anomaly detection based on convolutional recurrent autoencoder for IoT time series," IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 52, No. 1, pp. 112-122, 2020.
https://doi.org/10.1109/TSMC.2020.2968516

[12] Pearson, K., "LIII. On lines and planes of closest fit to systems of points in space," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Vol. 2, No. 11, pp. 559-572, 1901.
https://doi.org/10.1080/14786440109462720

[13] Salvador, S. and Chan, P., "Toward accurate dynamic time warping in linear time and space," Intelligent Data Analysis, Vol. 11, No. 5, pp. 561-580, 2007.
https://doi.org/10.3233/IDA-2007-11508

[14] Finance Data Reader, GitHub, Available online (accessed on 29 January 2024).
https://github.com/FinanceData/FinanceDataReader

[15] Wang, X., Yang, K., and Liu, T., "Stock price prediction based on morphological similarity clustering and hierarchical temporal memory," IEEE Access, Vol. 9, pp. 67241-67248, 2021.
https://doi.org/10.1109/ACCESS.2021.3077004

[16] Li, Y., et al., "Human activity recognition based on multienvironment sensor data," Information Fusion, Vol. 91, pp. 47-63, 2023.
https://doi.org/10.1016/j.inffus.2022.10.015

[17] Makridakis, S., Spiliotis, E., and Assimakopoulos, V., “The M4 Competition: 100,000 time series and 61 forecasting methods,” International Journal of Forecasting, Vol. 36, No. 1, pp. 54-74, 2020. https://doi.org/10.1016/j.ijforecast.2019.04.014

◐ 저 자 소 개 ◑

최 지 현(Ji-Hyun Choi)
2021년 한신대학교 정보통신학부(공학사)
2021년~현재 한신대학교 일반대학원 정보통신학과(석사과정)
관심분야 : 데이터 공학, 빅데이터 분석, 시계열 분석
E-mail : wlgus2391@hs.ac.kr

안 현(Hyun Ahn)
2011년 경기대학교 컴퓨터과학과(이학사)
2013년 경기대학교 컴퓨터과학과(이학석사)
2017년 경기대학교 컴퓨터과학과(이학박사)
2018년~2021년 경기대학교 컴퓨터공학부 연구교수
2021년~현재 한신대학교 AI·SW대학 조교수
관심분야 : 프로세스 마이닝, 데이터 공학
E-mail : hyunahn@hs.ac.kr