시차를 고려한 시계열 클러스터링 방법에 관한 연구

정재용*, 이주홍*, 송재원**
*인하대학교 전기컴퓨터공학과
**(주)밸류파인더스
eoehd4108@gmail.com, juhong@inha.ac.kr, jwsong@valuefinders.co.kr

A Study on Time Shifted Time Series Data Clustering

Jae-Yong Jeong*, Ju-Hong Lee*, Jae-Won Song**
*Dept. of Computer Engineering, Inha University
**ValueFinders Co., Ltd

요 약

데이터 클러스터링은 데이터의 숨겨진 패턴을 찾아낸다. 시계열 데이터에서 시차가 존재하는 데이터를 클러스터링하는 것은 데이터의 미래 패턴을 찾아내기 위해서 사용한다. 데이터 클러스터링을 수행하기 위한 여러 가지 Metric이 존재하지만, 시계열 데이터의 노이즈로 인해서 클러스터링을 수행하는 Metric을 설정하는데 제약이 존재한다. 본 논문은 기존 시계열 데이터가 가지고 있는 노이즈를 PIP 기법을 사용하여 제거하고, 노이즈가 없는 시계열 데이터를 클러스터링하기 위한 효율적인 새로운 Metric을 제안한다.

1. 서론

시계열 데이터란 연속적인 시간에 걸쳐 측정된 순차적인 데이터 집합이다[1]. 시계열 데이터의 종류로는 금융 데이터, 음성 데이터 등이 있다. 시계열 데이터에서 시차란 두 시계열 데이터의 관측지점 사이의 거리를 의미한다. 시계열 분석을 위한 방법으로, 시계열 데이터의 패턴을 찾거나 데이터의 유사한 그룹을 찾아내는 시계열 클러스터링 방법이 있다[2,3]. 본 논문은 다른 시간의 시계열 데이터들의 관계를 분석하기 위해 시계열에 시차를 포함하여 클러스터링을 수행하여 데이터들의 패턴을 파악한다. 클러스터링에서는 데이터의 유사도를 비교하기 위한 기준이 필요하다. 일반적으로 사용되는 유사도 Metric으로 거리, 상관관계 등이 있다. 시계열 데이터 클러스터링에서는 데이터에 포함된 노이즈로 인해서 클러스터링에 사용할 수 있는 Metric이 제한된다. 따라서 시계열 데이터 클러스터링에서 적절한 유사도 Metric을 찾는 것은 중요하다. 시계열 데이터의 노이즈 문제를 극복하기 위하여서 PIP(Perceptually Important Points) 알고리즘으로 데이터의 중요한 지점을 찾아내고, 나머지 데이터를 제거함으로 기존 시계열 데이터가 가지고 있는 노이즈를 제거한다. 노이즈가 제거된 시계열 데이터에서 기존의 Metric이 찾지 못하는 패턴을 찾아내기 위하여 새로운 유사도 Metric을 제안한다.

2. THE PROPOSED METRIC

2.1 Problem Statement

시계열 데이터 클러스터링을 위한 시계열 데이터를 정의한다. 시계열 데이터 $Y_{t:d}^{k}$는 k번째 시계열 데이터의 관측 시점 t에서 d까지 관측된 시계열 데이터를 나타낸다. 시계열 데이터를 다음과 같이 정의한다.

$$Y_{t:d}^{k} = \left\{ y_t^k, y_{t+1}^k, \cdots, y_d^k \right\}$$

데이터의 PIP를 찾는 보편적인 방법은 데이터의 양 끝 지점을 기준으로 만든 직선과 가장 멀리 떨어져 있는 데이터를 PIP로 정한다. 같은 과정을 반복하여서 다수의 PIP 찾아낼 수 있다. PIP를 찾는 방법은 데이터의 종류에 따라 다른 방법을 사용하다. 금융 데이터와 같이 데이터의 증감에 민감한 데이터의 경우는 PIP 변화의 증가와 감소가 반복되는 모양을 보

장하는 Zigzag-PIP 기법을 사용한다[4]. 본 논문에서는 데이터의 변동 비율이 임계치를 넘어가는 데이터를 PIP로 지정하였고, [5]에서 제공하는 ZigZag 패키지를 사용하여 구현하였다. 찾아낸 PIP를 제외한 나머지 데이터를 노이즈로 간주한다. 시계열 데이터 클러스터링을 수행하기 위해서는 모든 데이터가 동일 시점이어야 하는데, 시차가 존재하는 시계열 데이터의 시점은 서로 다르다. 이러한 이유로 시점이 다른 데이터들을 기준 시점으로 이동시킨다. 그래서 클러스터에 존재하는 모든 데이터는 모두 동일 시점 데이터로 표현되지만, 실제 각 데이터는 다른 시간 정보를 가진다.

2.1 클러스터링 Metric 제안

시계열 데이터 $Y^k_{t:d}$의 PIP 집합을 Z^k 이라고 정의한다. 두 개의 시계열 데이터 $Y^{k1}_{t:d}$, $Y^{k2}_{t:d}$ 의 유사도를 측정하는 Metric을 다음과 같이 유도한다. 먼저 두 데이터의 PIP를 공유하는 Z_{list}를 만든다.

$$Z_{list} = Z^{k1} \cup Z^{k2}$$

PIP 구간으로 데이터의 기울기 d^k_i을 다음과 같이 정의한다. (이때, $i \in Z_{list}$ 이다)

$$d^k_i = \frac{y_{z_i} - y_{z_{i-1}}}{z_i - z_{i-1}}$$

두 데이터 사이의 기울기의 유사도 함수 $SP(d^k_i, d^{k2}_i)$를 다음과 같이 정의한다.

$$SP(d^k_i, d^{k2}_i) = e^{-(d^k_i - d^{k2}_i)^2}$$

PIP와 실제 데이터 간의 오차를 기준으로 페널티 함수 $PN(y^k_i)$가 다음과 같이 정의한다.

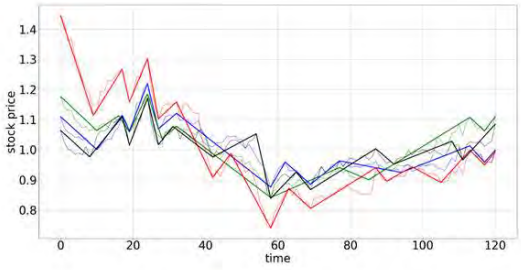
$$PN(y^k_i) = \sqrt{\frac{1}{(z_i - z_{i-1} - 2)} \sum_{i=z_{i-1}+1}^{z_i-1} (y^k_i - (\frac{y^k_{z_i} - y^k_{z_{i-1}}}{Z_i - Z_{i-1}} \times (i - Z_i) + y^k_{z_i})^2}$$

페널티 함수와 기울기 유사도 함수를 조합하여 다음과 같이 유사도를 정의한다.

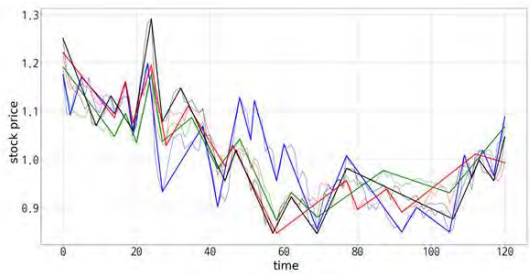
$$similarity(Y^{k1}_{t:d}, Y^{k2}_{t:d}) = \frac{\sum (\frac{SP(d^{k1}_i, d^{k2}_i)}{(PN(Y^{k1}_i) + PN(Y^{k2}_i))} \times (Z_i - Z_{i-1}))}{\sum Z_i - Z_{i-1}}$$

3. 실험

제안한 Metric을 평가하기 위해 거리기반의 클러스터링 방법인 K-Means 모델[6]과 제안한 Metric을 사용한 클러스터링 모델의 결과를 비교한다. 클러스터링 데이터는 1185개의 종목, 720일 기간의 주가 데이터를 평균이 1이 되도록 정규화하여, 단위 기간인 120일로 자르고 PIP를 적용하여 노이즈를 제거한 데이터를 사용하였다. 클러스터링의 조건은 클러스터 36개, 반복횟수 5를 기준으로 한다. 실험결과는 아래의 그림들과 같다. 그래프의 선은 동일 클러스터에 존재하는 종목별 주식데이터를 의미한다. 그래프의 굵은 선은 PIP를 사용하여 노이즈를 제거한 주식데이터이고, 얇은 선은 실제 주식데이터이다. 그림 1에서 (a)는 데이터 사이의 거리가 크더라도 데이터의 증감하는 추세가 비슷하게 유지되지만, (b)에서는 데이터 사이에 거리가 가까워도 데이터의 추세가 잘 반영되지 않음을 볼 수 있다. 이를 통하여 제안한 Metric이 거리기반의 클러스터링 방법 보다 데이터의 추세를 잘 반영함을 볼 수 있다.



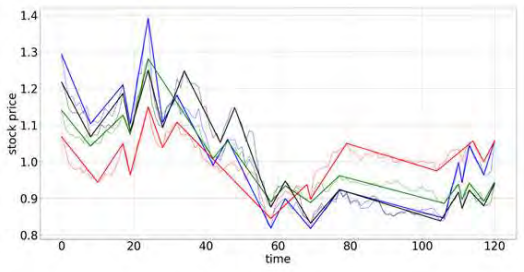
(a) 제안한 Metric 클러스터링



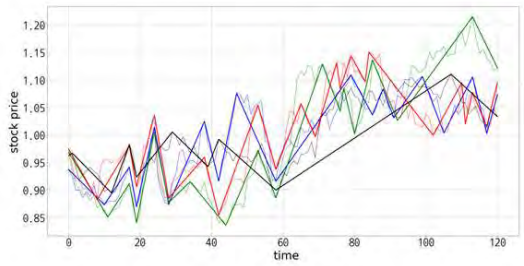
(b) K-means 클러스터링

(그림 1) 제안한 Metric과 K-means 클러스터링 비교

그림2에서 (a)와 (b) 모두 데이터의 추세를 반영하지만 (a)의 데이터 사이의 거리가 (b)에서의 데이터 사이의 거리보다 상대적으로 큰 것을 확인할 수 있다. 거리기반의 클러스터링 방법에서는 추세를 찾더라도 거리의 영향을 받지만, 제안한 Metric을 사용한 클러스터링은 거리의 제약을 받지 않고 데이터의 추세를 찾아낼 수 있음을 보여준다.



(a) 제안한 Metric 클러스터링



(b) K-means 클러스터링

(그림 2) 제안한 Metric과 K-means 클러스터링 비교

4. 결론

클러스터링에서 유사도를 정의하기 위한 여러 Metric이 존재하지만, 시계열 데이터에 적용하기 적절하지 않은 경우가 존재한다. 따라서 시계열 데이터 클러스터링을 위한 유사도 Metric을 제안하였다. PIP의 기울기 차이를 사용하여 기울기 유사도 함수를 정의하고, PIP가 반영하지 못한 실제 데이터 정보를 페널티 함수를 정의하여 반영한다. 기울기 유사도 함수와 페널티 함수를 조합하여 유사도 Metric을 정의한다. Metric을 평가하기 위한 실험에서는 주식데이터를 사용하고, 시차를 적용하여 다른 시점을 가진 시계열들의 패턴을 찾아낸다. 실험결과인 그림1과 그림2는 제안한 Metric을 사용한 클러스터링 모델과 K-Means 모델의 클러스터링 결과를 비교한 그림이다. 제안한 Metric을 사용한 클러스터링 모델이 K-Means 모델보다 데이터의 추세를 더 잘 반영함을 통하여 제안한 Metric의 성능을 검증하였다.

5. Acknowledgement

이 논문은 2019년도 정부(과학기술정보통신부) 의 재원으로 한국연구재단의 기초연구사업(과제번호: 2019R1F1A1062094)과 정보통신기획평가원의 지원(과제번호: 2019-0-01124)을 받아 수행된 연구임

참고문헌

[1] Adhikari, Ratnadip, and Ramesh K. Agrawal. “An introductory study on time series modeling a nd forecasting.” arXiv preprint arXiv:1302.6613 201 3.

[2] Roelofsen, Pjotr. “Time series clustering.” Vrij e Universiteit Amsterdam, Amsterdam, 2018.

[3] Pena, Daniel. “A course in time series analysi s.” Wiley-Interscience , 2011.

[4] Phetchanchai, Chawalsak, et al. “Index financia l time series based on zigzag-perceptually importa nt points.” Journal of Computer Science. 2010.

[5] “zigzag”, https://pypi.org/, last modified Jul 18, 2017, accessed Mar 23, 2020, https://pypi.org/proje ct/ZigZag/

[6] Arthur, David, and Sergei Vassilvitskii. “k-me ans++: The advantages of careful seeding.” Stanfo rd, 2006.