K-means 클러스터링 알고리즘을 이용한 financial transfer entropy 시각화 연구

김진규, 윤성로
고려대학교 전자전기공학과
e-mail: sryoon@korea.ac.kr (corresponding author)

A study on the visualization of financial transfer entropy by the k-means clustering algorithm

Jinkyu Kim, Sungroh Yoon
Dept of Electrical Engineering, Korea University

요 약

최근 Transfer entropy 이론을 주가지수 데이터에 적용하여 각 국가 간 상호 주고받는 정보의 방향성을 분석하고자 하는 연구가 진행되었다. 하지만 이렇게 각 국가 간의 정보 이동 관계를 고려하는 것도 중요하지만, 나아가 비슷한 방향성을 갖는 국가 군을 찾고 이를 분석하는 것 또한 중요한 연구이다. 기존의 연구 결과는 각 국가 간의 Entropy만 계산한 이차원 구조로 이 같은 경향성을 파악하기가 쉽지 않았다. 따라서 본 연구에서는 이 경향성을 쉽게 찾기 위해 k-means 클러스터링 알고리즘을 적용한 시각화 방법을 제안하고자 한다.

1. 서론

특정 두 국가 사이에서 서로 주고받는 정보와 그 영향을 측정하기 위해 correlation을 포함한 다양한 방법들이 제안되었다. 하지만 제안된 대부분의 방법들은 주고받는 정보의 절대적인 양을 측정하는 것에 중점을 두고 있을 뿐 국가 간 정보의 방향을 분석하기에는 쉽지 않았다.

이에 정보의 방향을 측정할 수 있는 Transfer entropy 이론을 각 국가의 주가지수 데이터에 적용하여 각 국가 간의 정보의 흐름을 파악하고 정보의 근원이 되는 국가를 찾고자 하는 연구가 제안되었다 [1]. 이 연구를 통해 한 국가의 주가지수가 타 국가로부터 받는 영향력을 계산해 낼 수 있었으며, 방향을 갖는 네트워크를 구축함으로써 정보의 흐름 또한 파악할 수 있었다. 여기서 Transfer entropy 이론은 집단 사이에 주고받는 정보의 절대적인 양 뿐만 아니라 정보의 방향성까지 파악하기 위해 제안된 이론이다 [2]. 방향성을 파악할 수 있다는 장점으로 현재 생명공학 분야, 금융 분야, 공학 분야 등 다양한 분야에 적용되고 있다.

하지만 이 연구에서는 단위 국가의 주가지수와 타 국가의 주가지수 간의 관계에 대해서만 초점을 맞춰 분석을 진행하였으며, 서로 비슷한 방향성을 가지는 국가들에 대한 고려는 하지 못했다. 또한, Transfer entropy 분석 결과의 특성상 결과가 특정 두 국가 간의 entropy 값으로 도출됨으로, 결과만 분석해서는 서로 비슷한 경향을 갖는 국가들을 엮어주는 것이 쉽지 않았다. 일례로 그림 1은 Financial 데이터를 적용하여 각 국가 간 엔트로피의 양을 도식화한 그림인데, 이 그림에서 볼 수 있듯이 결과 그림만으로 서로 비슷한 경향을 갖는 국가를 찾아내는 것이 힘들다는 것을 알 수 있다.

이에 본 논문에서는 기존에 연구가 진행된 2차원 클러스터링 알고리즘 개념을 Transfer entropy 분석에 접목하여 2차원 결과에서 비슷한 경향성을 보이는 국가들을 시각적으로 부각시킬 수 있는 방법을 제안하고자 한다.

2. 방법

2.1 실험데이터 및 데이터 전처리

실험을 위해 총 30개국[1)]의 주가지수를 수집하여 분석하였으며, 수집한 데이터는 2005년 10월 21일부터 2009년 3월 12일까지의 각 국가의 주가시가이다. 수집한 데이터는 참고문헌 [1]에서 언급한 것과 같은 전처리과정을 적용하였다. 간단히 요약하면, 먼저 각 데이터에 로그를 취한 뒤 일별 차이 값을 계산하였으며, 그 결과 값을 다음과 같이 세 가지 상태로 분류한 뒤, 알고리즘을 적용하였다; $A(n)=0$ (if $x(n) \le -d$), $A(n)=1$ (if $-d<x(n)<d$), $A(n)=2$ (if $x(n) \ge d$).

1) 분석한 30개국 목록: 아르헨티나, 오스트리아, 브라질, 캐나다, 칠레, 중국, 덴마크, 핀란드, 프랑스, 영국, 독일, 그리스, 네델란드, 홍콩, 헝가리, 인도, 인도네시아, 이탈리아, 일본, 한국, 말레이시아, 멕시코, 뉴질랜드, 러시아, 싱가폴, 스페인, 스위스, 미국, 베네수엘라, 베트남

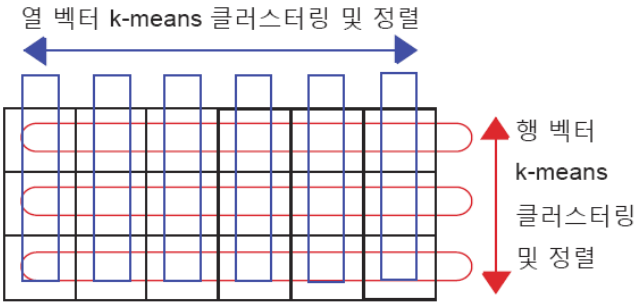

그림 1 k-means 클러스터링 적용 방법

2.2 Transfer entropy 이론 적용

Transfer entropy 이론은 두 가지 Discrete time series 데이터 간의 방향성을 측정하는데 주로 적용되며, 식 1과 같이 정의된다 [2].

$$T_{J\to I} = \sum p(i_{t+1}, i_t^k, j_t^l)\log\frac{p(i_{t+1}|i_t^k, j_t^l)}{p(i_{t+1}|i_t^k)} \quad (\text{식 } 1)$$

본 연구에서는 k와 l의 값으로 1을 사용했으며, 이는 데이터 분석에서 특정 시간 t와 t-1에서의 값만 고려하는 것을 의미한다.

2.3 K-means 클러스터링 알고리즘

본 논문에서 적용한 k-means 클러스터링 알고리즘은 Unsupervised learning의 대표적인 한 종류로, 이 알고리즘의 핵심은 주어진 데이터에서 사용자가 지정한 k 개에 해당하는 클러스터의 중심을 찾아내는 것이다.

본 논문에서는 최적화 된 클러스터의 개수를 찾기 위해 k의 값을 1부터 총 분석 국가의 수인 30까지 증가시키면서 실루엣 값을 분석하였고, 가장 작은 값을 가지는 k의 값을 클러스터의 수로 사용하였다. 또한, Euclidean 방법을 사용하여 데이터 간 거리를 측정했으며, 각 10번의 반복을 통해 알고리즘을 학습시켰다.

한편, 2차원 데이터에 알고리즘을 적용하기위해 그림 1과 같이 두 단계로 나눠서 분석하였다. 먼저, 행벡터를 추출하여 알고리즘을 적용한 뒤 클러스터 별로 정렬하였다. 이 후, 열벡터에 대해서도 같은 방식으로 적용한 뒤 정렬하여 최종 결과를 도출하였다.

3. 결과

앞서 설명한 방법을 적용하여 결과를 분석하였으며, 그림 2에서 그림 5까지 결과를 정리하였다. 그림 2의 경우, Transfer entropy(TE) 이론을 적용한 후 별도의 처리를 하지 않은 결과이다. 실험에 사용된 국가의 수는 30개국으로 국가 상호간의 entropy를 계산해 30×30 matrix로 결과를 정리하였다. x축의 국가는 정보를 제공하는 국가이며 y축은 정보를 받는 국가를 의미한다. 따라서 matrix의 각 값은 x축의 국가가 y축의 국가에게 주는 정보의 양이며, 짙은 색일수록 많은 양의 정보를 주고받고 있다는 것을

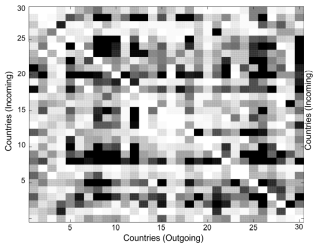

그림 2 TE (적용 전)

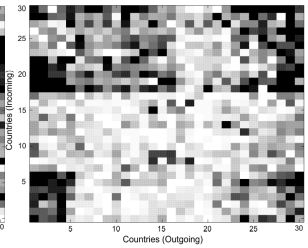

그림 3 TE (적용 후)

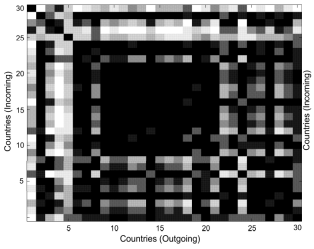

그림 4 correlation(적용 전)

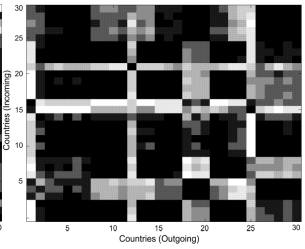

그림 5 correlation(적용 후)

의미한다. 그림 2의 경우, 각 국가 간의 정보의 양을 파악하는 데는 용이하지만 전체적인 국가 간 경향성을 파악하기에는 한계를 가지는 것을 확인할 수 있다.

그림 3의 경우, k-means 클러스터링 알고리즘을 적용하여 정렬한 그림이다. 그림 3의 경우, 그림 2와는 달리 비슷한 경향을 가지는 국가들을 정렬함으로써 한층 국가 간 경향성을 파악하기가 용이해 짐을 알 수 있다.

추가적으로 Transfer entropy 이외에 correlation을 이용하여 분석한 결과에도 k-means 클러스터링 알고리즘을 적용하였고, 분석한 결과를 그림 4-그림 5에 정리하였다.

그림 4의 경우, 각 국가 간 시간의 흐름을 고려하지 않고 correlation 값을 계산한 결과이며, 그림 5의 경우 k-means 클러스터링 알고리즘을 적용한 결과이다. 이 결과에서 볼 수 있듯이 k-means 클러스터링 알고리즘을 적용함으로써 비슷한 경향성을 갖는 국가 군으로 쉽게 분석할 수 있는 것을 확인할 수 있다.

4. 결론

본 연구를 통해 기존에 각 국가 별로 주고받는 정보의 양을 분석하는 것에서 나아가 비슷한 경향성을 가지는 국가군을 k-means 알고리즘을 적용하여 사람의 눈으로 쉽게 분석할 수 있는 방법을 제안하였다. 즉, 기존에 Transfer entropy를 적용한 뒤 별도의 처리를 하지 않고 얻은 결과는 경향성을 파악하기가 쉽지 않은 반면, 제안한 방법을 통해 국가 간의 경향성을 쉽게 파악할 수 있었다.

참고문헌

[1] Kwon O. and Yang J.-S., "Information flow between stock indices," EPL, 82 (2008) 68003.

[2] Schreiber T., Phys. Rev. Lett., 85 (2000) 461.