엔트로피 기반 인과관계 네트워크의 모듈성을 활용한 상품 선물 시장의 EDaR 변동 예측 모형 개발

최인수 [1], 김우창 [1]
[1] 한국과학기술원 산업및시스템공학과
jl.cheivly@kaist.ac.kr, wkim@kaist.ac.kr

Developing an Entropic Drawdown-at-Risk (EDaR) Fluctuation Forecasting Model for Commodity Futures Market Using Entropy-Based Dependency and Causality Network Modularity

Insu Choi[1], Woo Chang Kim[1]
[1]Dept. of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST)

요 약

본 연구에서는 전이 엔트로피 개념을 활용하여 주요 상품 선물의 하방 리스크 지수의 정보 흐름을 바탕으로 한 인과관계 네트워크를 구성하였다. 그리고 구성된 네트워크를 활용하여 금융 시장을 분석하였으며, 또한 정보 흐름의 존재 여부를 바탕으로 상품 선물의 하방 리스크 지수의 예측력이 개선될 수 있는지 확인하고자 하였다. 이를 위하여 정보 불확실성의 감소량을 측정하는 전이 엔트로피를 인과관계의 측정 지표로 상정하였으며, 전이 엔트로피 측정 시 발생할 수 있는 유한크기 효과(finite size effect)를 조정하는 데 있어서 효과적인 지표인 효율적 전이 엔트로피를 활용하여 정보 흐름 네트워크를 구성하였으며 이를 이용하여 금융 지수 간의 인과관계를 분석하고 EDaR 의 등락 예측에 활용하였다. 그 결과, 금융 시장 지수를 효율적 전이 엔트로피를 이용한 인과관계 네트워크를 활용하여 금융 시장의 복잡계 네트워크 분석이 가능함을 확인하였고, 구성된 네트워크를 활용하여 국내 금융 시장 등락 예측에 있어 더 적은 데이터 열을 활용하여 거의 유사한 예측 결과를 냄으로써 상품 선물 시장 관련 예측의 데이터 열 선택에 활용할 수 있음을 확인하였다.

1. 서론

상품 선물(commodity futures markets)은 2000 년대 초반 이후 뚜렷한 거래량의 증가세를 보이며 많은 관심을 받고 있다. 2008 년 미국 상품 선물거래위원회(Commodity Futures Trading Commission, CFTC, 2008)의 보고서에 따르면 2008 년 6 월 기준 상품지수 순투자액은 2,000 억 달러로 2003 년보다 10 배 이상 증가했으며 2009 년에는 약 2,500 억 달러로 증가한 것으로 알려져 있으며 (Irwin and Sanders, 2011) 이러한 추세는 2010 년대 후반에도 지속적으로 나타나고 있다. 기관투자자, 인덱스펀드, 국부펀드 및 ETF(상장지수펀드), ETN(상장지수증권) 및 이와 유사한 상품을 보유하고 있는 개인투자자가 미국 시장에서 거래되는 주요 상품 선물시장의 투자자들이며 향후 이러한 투자 비중이 점점 증가할 것으로 보인다.

기존 금융 상품 외에 투자 상품을 다각화하는 용도로 활용되는 전통적인 역할 외에도, 상품 선물은 특히 주식 시장의 금융 위기와 침체 기간 동안 포트폴리오의 위험 분산에 있어 높은 잠재력을 보이는 새로운 자산군으로 새롭게 떠오르고 있다. Shannon (1948)이 제시한 정보 엔트로피 개념을 이용한 금융 시장 환경 분석의 경우에는 금융 시장 역시 정보의 집합체라는 관점으로부터 시작되며 이를 통해 비선형적 의존성(하나의 확률변수가 다른 하나의 확률변수에 대해 제공하는 정보의 양), 비선형적 인과성(정보의 불확실성 감소량) 등을 측정할 수 있는 다양한 설명 가능한 개념이 있다는 점에서 대표적인 경제물리학적 접근 방법으로 꼽힌다. 정보 엔트로피의 개념이 경제학 및 금융 분야에 처음 등장한 것은 Georgescu-

Roegen (1971)의 저서 <The Entropy Law and the Economic Process>이며, 이 당시는 열역학적 개념을 통해 경제학을 설명하는 요소로 활용하기 시작하였으며, 이후 금융 시장에서도 수익률 등의 경험적 분포가 일반적으로 정규성을 따르지 않는다는 실증적 결과들이 나옴에 따라 이러한 정규성 가정에 대한 대체재로 금융 분야에서 금융 시장 분석에 활용되기 시작하였다. 이러한 정보 엔트로피는 데이터의 성질을 전제로 하지 않는다는 특징이 있으며 이에 본 연구에서는 이러한 엔트로피의 개념을 활용하여 인과관계 네트워크를 구성하고자 한다.

2. 연구 방법론

2.1 전이 엔트로피 (Transfer Entropy, TE)

전이 엔트로피는 Schreiber (2000)에 의해 개발이 되었으며, 컴퓨터 과학, 뇌 과학, 사회 관계 분석, 인과관계 분석 및 응용통계학 등 여러 분야에서 주로 활용되었다. 금융 분야에서도 역시 시장 분석에 있어 2010 년대 초중반부터 주로 활용되어 왔다. TE 의 식은 다음과 같다.

$$TE_{Y\to X}^{(k,l)}(t) = I(x_{t+1}, y_t^{(l)} | x_t^{(k)}) = H(X_{t+1}|X_t, \dots, X_{t-k+1}) - H(X_{t+1}|X_t, \dots, X_{t-k+1}, Y_t, \dots, Y_{t-l+1})$$
$$= \sum_i p\left(x_{t+1}, x_t^{(k)}, y_t^{(l)}\right) \log_2 p(x_{t+1}|x_t^{(k)}, y_t^{(l)}) - \sum_i p\left(x_{t+1}, x_t^{(k)}, y_t^{(l)}\right) \log_2 p(x_{t+1}|x_t^{(k)})$$
$$= \sum_i p(x_{t+1}, x_t^{(k)}, y_t^{(l)}) \log_2 \frac{p(x_{t+1}|x_t^{(k)}, y_t^{(l)})}{p(x_{t+1}|x_t^{(k)})}$$

$$\textbf{(1)}\ TE_{Y\to X}(1,1) = H(X_{t+1}|X_t) - H(X_{t+1}|X_t, Y_t) = \sum_i p(x_{t+1}, x_t, y_t) \log_2 \frac{p(x_{t+1}, x_t, y_t) p(x_t)}{p(x_{t+1}, x_t) p(x_t, y_t)}$$

$$\textbf{(2)}\ ETE_{Y\to X} = \frac{TE_{Y\to X}(k,l) - \frac{1}{M}\sum_{i=1}^{M} TE_{Y_{(t)}\to X}(k,l)}{\sigma_M}$$

TE 의 경우 유한 표본 효과(finite size effect)로 인해 그 값이 과소평가 또는 과대평가될 수가 있어 이러한 문제점을 해결하기 위해 (2)와 같은 형태의 효율적 이전 엔트로피(effective transfer entropy, ETE) 개념이 Boba et al. (2015)에 의해 제시되었으며, 본 연구에서는 해당 방법론을 활용

2.2 네트워크 분석

네트워크의 군집을 생성하기 위하여 다음과 같은 Newman(2006)과 Dugué and Perez (2015)의 무향 네트워크의 유향 네트워크에 대한 Louvain 알고리즘에 관한 연구를 활용하였다. 해당 연구에서는 유향 네트워크의 군집화를 위하여 다음과 같은 네트워크의 모듈성(modularity)을 최대화하는 방식으로 활용

이 때 모듈성(modularity)은 네트워크 내에서 상대적으로 밀접한 관계를 가지고 있는 하위 집단을 찾기 위해 주로 사용되는 척도로, 모듈성이 크다는 것은 집단 내 정점(상품 선물 지수)들 사이의 관계가 집단 간 상품 선물의 하방 리스크 사이의 관계에 비해 강한 것을 의미한다. 다시 말해, 하위 그룹이 명확하게 분류될수록 모듈성의 값이 큼을 의미한다.

드로우다운(Drawdown)은 특정 기간 동안 발생한 최고점에서 최저점까지의 손실금액으로, 주로 하방리스크(downside risk)를 측정하는 대표적인 지표이다.

Cajas (2021)에 의해 제시된 EDaR 은 특정 기간에서 수준을 넘어서는 드로우다운의 평균에 대한 상한(upper bound)임이 알려져 있으며, EDaR 의 식은 다음과 같다.

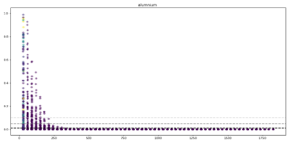
$$EDaR_\alpha(X) = \inf_{z>0}\left\{ z\ln\left(\frac{M_{DD(X)}(z^{-1})}{\alpha}\right)\right\}$$

$$DD(X, j) = \max_{t\in(0,j)}\left(\sum_{i=0}^{t} X_i\right) - \sum_{i=0}^{j} X_i$$

2.3 기계 학습

기계 학습 모형으로는 Chen and Guestrin(2016), LightGBM 은 Ke et al. (2017), Prokhorenkova et al.(2018)이 개발한 XGBoost 를 활용하였으며 각 데이터열에 대한 영향도 해석을 위해 SHAP 를 사용하였다. Lundberg and Lee (2017)이 제시한 SHAP 값은 조건부 기대 함수의 섀플리 가치를 기반으로 제시되는 값이다. SHAP 값은 각 특징(Feature)이 학습된 모형에 대해 어느 정도 기여하는 지를 측정한다.

3. 실험 결과

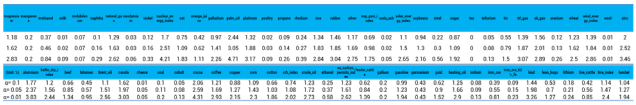


<그림 1> 알루미늄 데이터 세트의 정규성 p-value 값 분포

상품 선물 시장에서 거래되고 있는 주요 상품 선물 38 개 데이터에 대해서 2017 - 2021 년의 데이터를 취합하였으며, 2017 - 2020 년 데이터를 학습 세트로, 2021 년 데이터를 테스트 세트로 설정하였다.

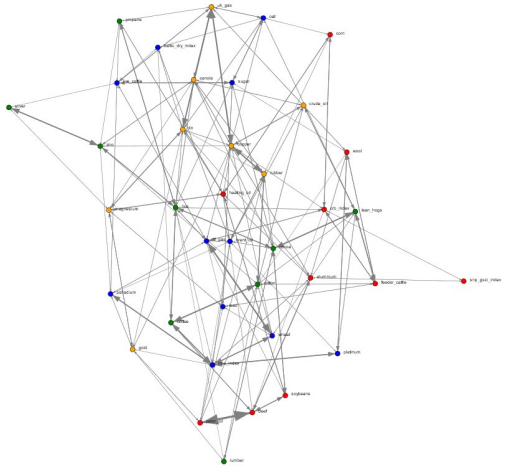
이 기간 내 최소 월 단위부터 최대 5 개년 (2017 - 2021) 데이터까지 각 데이터 별로 1,830 개의 데이터 세트에 대해서 7 개의 정규성 검정(Shapiro-Wilk Test (W), D'Agostino K-squared Test (K-Squared), Lilliefors Test (T), Jarque-Bera Test (JB), Kolmogorov- Smirnov Test (KS), Anderson-Darling Test (A-Squared), Cramér- von Mises Test (U))을 실시한 결과 정규성을 만족하는 비율은 0% - 5% 사이이며 기간이 길어질수록 이 비율이 매우 감소함에 따라 사실상 경험적 분포에서 정규성을 만족하는 경우가 거의 없다고 해석할 수 있다.

이에 이러한 정규적이지 않은 EDaR 변화율 데이터의 통계적 성질을 본 연구에서는 통계적 성질에 대한 전제를 갖지 않는 전이 엔트로피를 사용하고자 하는 정량적 근거로 활용하였다.

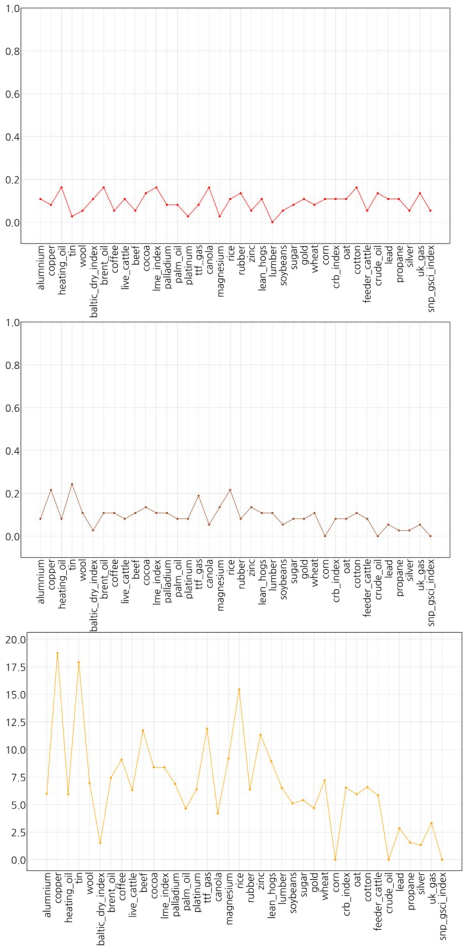


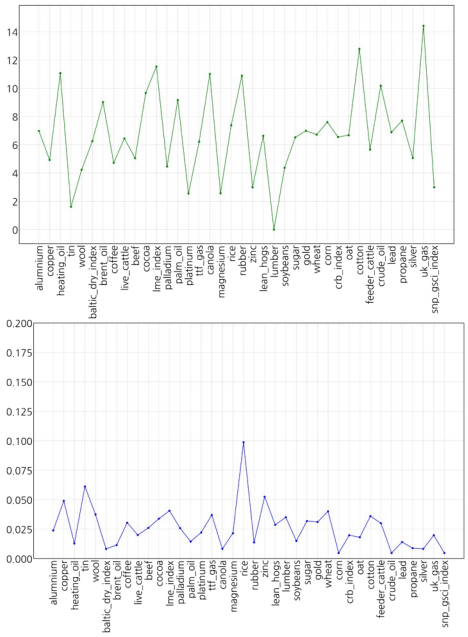
<표 1> 데이터 세트의 정규성 만족 비율

이러한 통계적 근거를 바탕으로 효율적 전이 엔트로피를 계산하고, 유의 수준 0.01 하에서 구성된 EDaR 의 정보 흐름 네트워크는 다음과 같다.



<그림 2> 상품 시장 선물 EDaR 네트워크





<그림 3> 네트워크 Out-Degree, In-Degree, Out-Strength, In-Strength, PageRank 값

XGBoost, LightGBM, CatBoost 3 개의 모형을 이용한 소프트 보팅 결과에 대한 비교 결과는 다음과 같음

전체 실험의 예측에 대한 SHAP 값의 활용 순위 평균과 네트워크의 주요 중심성 지표와는 다음과 같이 약한 양의 상관관계를 지닌다.

<표 2> 중심성 지표 분석 결과

In-Degree	Out-Degree	Out-Strength	In-Strength	PageRank
0.1817	0.0413	0.1959	0.1586	0.2320

38 개의 실험 결과에 대한 성과 평가 지표를 독립 t-검정을 이용하여 비교한 결과 $\alpha = 0.05$ 수준에서 유의미한 변동이 없음

이는 네트워크의 구조를 통해서 데이터의 약 ¼에 해당하는 데이터만 활용하여도 상품 선물 시장 전체 데이터를 사용하는 것과 유사한 효과로 해석할 수 있다. 또한, 우측 <표 5>에서 확인할 수 있듯 평균적으로 정확도 등 일부 항목 소폭 개선 효과 역시 있다.

<표 3> 원 예측 결과와 클러스터를 이용한 예측 결과의 성과 지표에 대한 대응 표본 t-검정 결과

Paired T-test (Two-sided)	T-Statistic	p-value
Accuracy	0.7078	0.4834
Balanced Accuracy	-1.5812	0.1221
Cohen-Kappa Score	-1.4238	0.1627
Precision Score	-0.9692	0.3386
Recall Score	0.7078	0.4834
F1 Score	-1.7103	0.0953
F-Beta Score (0.5)	-1.7586	0.0867
F-Beta Score (1)	-1.7104	0.0953
F-Beta Score (2)	-0.6201	0.5389
Hamming Loss	-0.708	0.4834

Feature	mean(|SHAP|)	Rank
crude_oil	0.3778	1
brent_oil	0.3290	2
wool	0.3032	3
lumber	0.1916	4
copper	0.1882	5
corn	0.1773	6
ttf_gas	0.1770	7
rice	0.1641	8
heating_oil	0.1596	9
tin	0.1595	10
silver	0.1585	11
palladium	0.1584	12
palm_oil	0.1558	13
platinum	0.1540	14
canola	0.1506	15
feeder_cattle	0.1470	16
magnesium	0.1455	17
gold	0.1442	18
uk_gas	0.1405	19
baltic_dry_index	0.1394	20
oat	0.1386	21
lme_index	0.1347	22
rubber	0.1319	23
soybeans	0.1279	24
sugar	0.1252	25
beef	0.1239	26
cotton	0.1229	27
coffee	0.1219	28
alumnium	0.1208	29
wheat	0.1201	30
zinc	0.1195	31
propane	0.1189	32
live_cattle	0.1138	33
lean_hogs	0.1137	34
lead	0.1116	35
cocoa	0.1111	36
crb_index	0.0870	37
snp_gsci_index	0.0820	38

<표 4> SHAP 순위

Performance Measure	Cluster-Based	Original
True Negative	193	184
False Positive	40	48
False Negative	100	93
True Positive	32	39
Accuracy	0.6165	0.6132
Balanced Accuracy	0.5308	0.5388
Cohen-Kappa Score	0.0674	0.0830
Precision Score	0.5805	0.5860
Recall Score	0.6165	0.6132
F1 Score	0.5807	0.5896
F-Beta Score (0.5)	0.5745	0.5843
F-Beta Score (1)	0.5807	0.5896
F-Beta Score (2)	0.5988	0.6018
Hamming Loss	0.3835	0.3868

<표 5> 분석 결과

4. 결론

본 연구에서는 전이 엔트로피 개념을 활용하여 주요 상품 선물의 하방 리스크 지수의 정보 흐름을 바탕으로 한 인과관계 네트워크를 구성하였다. 그리고 구성된 네트워크를 활용하여 금융 시장을 분석하였으며, 또한 정보 흐름의 존재 여부를 바탕으로 상품 선물의 하방 리스크 지수의 예측력이 개선될 수 있는지 확인하였다.

이를 위하여 정보 불확실성의 감소량을 측정하는 전이 엔트로피를 인과관계의 측정 지표로 상정하였으며, 전이 엔트로피 측정 시 발생할 수 있는 유한크기 효과(finite size effect)를 조정하는 데 있어서 효과적인 지표인 효율적 전이 엔트로피를 활용하여 정보 흐름 네트워크를 구성하였으며 이를 이용하여 금융 지수 간의 인과관계를 분석하고 EDaR 의 등락 예측에 활용하고자 하였다.

그 결과, 금융 시장 지수의 하방 리스크의 상한을 효율적 전이 엔트로피를 이용한 인과관계 네트워크를 활용하여 금융 시장의 복잡계 네트워크 분석이 가능함을 확인하였고, 구성된 네트워크를 활용하여 국내 금융 시장 등락 예측에 있어 더 적은 데이터 열을 활용하여 거의 유사한 예측 결과를 냄으로써 상품 선물 시장 관련 예측의 데이터 열 선택에 활용할 수 있음을 확인하였다.

한계점의 경우 성능의 문제로 인하여 그래디언트 부스팅 알고리즘 기반의 기계 학습 방법론을 활용하여 예측 및 성과 측정을 진행하였으나 최신의 다양한 기계 학습 방법론을 추가적으로 활용하여 이것이 그래디언트 부스팅 알고리즘과 같은 특정 형태의 기계 학습 기반 예측 방법론에 국한되지 않는다는 것을 보임으로써 결과의 강건성을 높일 필요가 있다.

참고문헌

[1] Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.

[2] Georgescu–Roegen, N. (1971). The entropy law and the economic process. Harvard university press.

[3] Schreiber, T. (2000). Measuring information transfer. Physical review letters, 85(2), 461.

[4] Boba, P., Bollmann, D., Schoepe, D., Wester, N., Wiesel, J., & Hamacher, K. (2015). Efficient computation and statistical assessment of transfer entropy. Frontiers in Physics, 3, 10.

[5] Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-8582.

[6] Dugué, N., & Perez, A. (2015). Directed Louvain: maximizing modularity in directed networks (Doctoral dissertation, Université d'Orléans).

[7] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[8] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[9] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

[10] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[11] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.