히스토그램 기반 상호 정보량 지표를 활용한 전체 그래프 임베딩 기반의 수익률 예측

최인수 [1], 김우창 [2*]
[1]KAIST 산업및시스템공학과
[2]KAIST 산업및시스템공학과

jl.cheivly_@kaist.ac.kr, wkim_@kaist.ac.kr

Financial Asset Return Prediction via Whole-Graph Embedding Leveraging Histogram-Based Mutual Information

Insu Choi[1], Woo Chang Kim
[1]Dept. of Industrial and Systems Engineering, KAIST
[2]Dept. of Industrial and Systems Engineering, KAIST

요 약

본 논문에서는 정보 이론 기반 지표의 힘을 활용하여 전체 그래프 임베딩 방법론의 한 가지인 GL2vec 을 사용하여 임베딩을 생성하고, 이를 바탕으로 상장지수펀드 (ETF, Exchange Traded Fund) 수익률을 예측하는 모형을 생성하고자 하였다. 본 연구는 그래프 구조에 금융 데이터를 내장하고 고급 신경망 기술을 적용하여 예측 정확도를 향상시키는 데에 기여할 수 있음을 확인하였다.

1. 서론

본 연구는 금융 상품의 수익률 예측 개선을 목적으로 그래프 기반 방법론을 활용하고자 하였다. 여기서 그래프의 각 요소는 주요 금융 시장을 의미하며, 각 요소는 시장 지수, 두 요소 간의 관계는 다양한 이동 창 (20 일, 60 일, 120 일)의 히스토그램을 활용한 상호정보량을 통해 연결되었다. 이 방법을 통해 금융 시장의 복잡한 구조를 분석하는 데 GL2Vec 을 사용하였다. 결과적으로, 이 방식은 순환 신경망보다 더 정확한 예측을 제공하였다.

본 연구는 상품 선물의 복잡한 구조를 이해하고 예측하기 위해 네트워크 방식에 중점을 두었다. 기존의 금융 시장 예측은 주로 통계 모델을 기반으로 하였지만, 이러한 전통적인 방법에는 한계가 있다. 반면, 최근에는 신경망 기반의 모델이 주목받고 있어, 이를 활용하였다. 본 연구의 주된 목표는 금융 시장의 복잡한 구조를 네트워크로 표현하고, 이를 기반으로 예측 성능을 향상시키는 것이었다.

<표 1>의 데이터를 사용하여 상품 선물의 복잡한 상호 작용을 분석하였다. 상호정보량을 파악하기 위해 다양한 이동 창의 히스토그램을 활용하였다. 그리고, 다양한 시간 간격의 데이터를 통해 강건한 예측 모델을 구축하려 하였다. GL2vec 알고리즘을 사용하여 전체적인 그래프 구조를 동시에 분석하였다. 그 결과, 제안한 모델은 기존 방식보다 더 나은 예측 성능을 보였다.

2. 데이터 설명

본 연구의 분석 대상으로는 뉴욕증권거래소 (New York Stock Exchange Archipelago, NYSE Arca)에서 거래되는 섹터 ETF 를 설정하였으며, 그 결과 9 개의 스타일 / 크기 ETF 종가 데이터를 연구에 활용하였다. 활용된 데이터에 대한 설명은 <Table 1>과 같다:

Table 1. 연구 실험 대상

Full Name	Abbreviation	Tracking Index	Asset Class
Vanguard Large-Cap ETF	VV	CRSP US Large Cap Index	Equity
Vanguard Small-Cap ETF	VB	CRSP US Small Cap Index	Equity
Vanguard Mid-Cap ETF	VO	CRSP US Mid Cap Index	Equity
Vanguard Value ETF	VTV	CRSP US Large Cap Value Index	Equity
Vanguard Growth ETF	VUG	CRSP US Large Cap Growth Index	Equity
Vanguard Small-Cap Value ETF	VBR	CRSP US Small Cap Value Index	Equity
Vanguard Small-Cap Growth ETF	VBK	CRSP US Small Cap Growth Index	Equity
Vanguard Mid-Cap Value ETF	VOE	CRSP US Mid Cap Value Index	Equity
Vanguard Mid-Cap Growth ETF	VOT	CRSP US Mid Cap Growth Index	Equity

연구 데이터는 2013 년 1 월부터 2022 년 12 월까지의 10 개년 데이터를 활용하였다.

3. 상호 정보량

정보 이론에서 시스템 X 의 동작은 확률 분포 p(x)와 p(x)의 로그 값으로 정의되며, 이 개념에 기초하여, 정보 엔트로피는 다음과 같이 공식화할 수 있다 (Shannon, 1948):

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x,y) \quad (1)$$

여기서 p(x)는 데이터 X 의 확률 분포입니다. 엔트로피는 확률 분포의 불확실성을 측정하는 지표로 해석될 수 있습니다.

두 데이터 X 와 Y 가 있을 때, 조건부 엔트로피 (Conditional Entropy) X 가 주어졌을 때 Y 의 조건부 엔트로피는 다음과 같이 정의할 수 있다.

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x) \log_2 \frac{p(x,y)}{p(x)} \quad (2)$$

또한, 결합 엔트로피는 다음과 같이 설명할 수 있다:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x,y)$$

상호 정보량의 정의 상호 정보량은 두 확률 변수 X 와 Y 사이의 정보 공유 정도를 측정합니다. 수식으로 표현하면:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (4)$$

일반적으로, 실제 데이터에서는 연속된 값들을 이산화하여 히스토그램 형태로 나타냅니다. 데이터의 분포를 히스토그램으로 근사화한 후, 각 구간에서의 확률을 추정하여 위의 엔트로피 및 상호 정보량을 계산한다.

MI 를 계산할 때 히스토그램의 구간 수와 크기는 중요한 역할을 한다. 구간의 수와 크기를 변경함으로써 결과가 크게 달라질 수 있으므로, 이 부분은 신중하게 선택해야 한다. 또한 구간의 크기가 너무 크면 정보가 손실될 수 있으며, 너무 작으면 과적합의 위험이 있다. 이 중 본 연구에서는 Hacine-Gharbi et al. (2018)이 제시한 상호 정보량에 수치해석적으로 최적으로 알려진 구간의 값을 활용하였다.

MI 는 특히 변수 선택, 클러스터링, 이미지 등록, 데이터 마이닝 등 다양한 분야에서 응용된다. MI 가 높으면 두 변수가 서로 높은 정보를 공유하고 있음을 의미한다.

본 연구에서 상호정보량(mutual information, MI)을 사용한 주요 이유는 다음과 같이 요약할 수 있다.

상호정보량은 두 변수 간의 모든 종류의 종속성, 즉 선형 및 비선형 관계를 측정할 수 있다. 금융 시장 데이터는 종종 비선형 특성을 가지기 때문에, Pearson 상관계수와 같은 전통적인 선형 상관 측정 방법만 사용할 경우 데이터의 모든 관계를 완전히 포착할 수 없다. 상호정보량을 사용하면 이러한 비선형 관계도 포착할 수 있다. 이외에도 상호정보량은 두 변수가 독립적일 때 0 의 값을 가지며, 변수 간의 종속성이 증가할수록 값이 증가한다. 따라서 금융 시장에서 여러 변수 간의 복잡한 상호 작용과 의존성을 정량적으로 평가하는 데 적합하다.

또한, 상호정보량은 정보 이론의 개념에 기반하며, 두 변수의 공통 정보량을 측정한다. 이는 금융 시장에서 변수 간의 정보 공유 정도를 정확하게 측정하는 데 유용하다.

상호정보량은 비모수적인 방법이므로 다양한 데이터 분포와 구조에 대해 강건하다. 금융 데이터의 변동성과 불규칙성을 고려할 때, 이러한 강건성은 예측 모델의 안정성과 정확성을 높이는 데 중요하다. 마지막으로 금융 시장의 데이터는 매우 복잡하며 다양한 요소가 상호 작용한다. 상호정보량은 이러한 복잡한 데이터 구조와 상호 작용을 정보 이론에 입각하여 효과적으로 해석하고 분석하는 데 도움을 준다.

결론적으로, 상호정보량은 금융 시장의 복잡한 구조와 상호 작용을 효과적으로 분석하고 이해하는 데 필수적인 도구다. 이 연구에서는 상호정보량의 이러한 장점을 활용하여 금융 시장의 복잡한 상호 작용을 효과적으로 분석하고 예측하였다.

4. GL2Vec (Chen and Koga, 2019)

GL2vec 는 그래프 임베딩 방법 중 하나로, 그래프의 구조와 연결 특성을 함께 고려하여 임베딩을 생성한다. 기본 아이디어는 "선분 그래프"를 활용하는 것이다.

선분 그래프 L(G)는 원래의 그래프 G 에서 연결을 노드로, 공통된 노드를 공유하는 연결 쌍을 연결로 가지는 그래프이다. 수식으로 표현하면, G 의 연결 집합을 E 라 할 때, L(G)의 노드 집합은 E 가 된다.

GL2vec 의 주요 과정은 다음과 같다:

1. 주어진 그래프에서 선분 그래프 L(G)를 생성한다.
2. L(G)의 노드 (즉, 원래 그래프의 연결)에 대한 임베딩을 학습한다. 이 때 임베딩 방법으로는 다양한 그래프 임베딩 방법 (예: DeepWalk, Node2Vec 등)을 사용할 수 있다.
3. 학습된 임베딩을 원래 그래프의 연결 특성과 함께 통합하여 최종 임베딩을 생성한다.
4. 이 방법을 통해, 그래프의 구조적인 정보와 연결의 특성 정보를 모두 활용하여 더 풍부하고 정확한 임베딩을 생성할 수 있다. 본 연구에서는 8 차원의 임베딩을 통해 추출된 결과에 대해 예측을 실시하였다.

GL2Vec 는 전체 그래프의 구조를 고려하여 그래프를 벡터로 임베딩하는 것을 목표로 합니다. 따라서 개별 노드의 특성뿐만 아니라 그래프 전체의 구조와 패턴을 효과적으로 표현할 수 있다. 또한, GL2Vec 는

다양한 종류의 그래프, 예를 들면 방향성이 있는/없는 그래프, 가중치가 있는/없는 그래프 등에 모두 적용할 수 있으며 GL2Vec 는 그래프의 크기나 복잡도에 상관없이 임베딩을 빠르고 효율적으로 수행할 수 있다. 이외에도 GL2Vec 는 다양한 분야의 연구나 응용에 활용될 수 있으며, 특히 그래프 기반의 복잡한 데이터셋에서 그 유용성이 입증된 바 있다.

5. 실험 설명 및 결과

본 연구에서는 세 가지 다른 이동창 길이, 즉 20일, 60일, 그리고 120일에 대하여 GRU (Gated Recurrent Unit) (Chung et al., 2014) 모형을 활용하여 분석을 진행하였다. GRU 는 순환 신경망의 한 종류로서, 내부 게이트 메커니즘을 통해 정보의 흐름을 조절하여 장기 의존성 문제를 해결하는 데 특화되어 있다. 본 연구에서는 가장 기본적인 단일 레이어 GRU 모델을 사용하였다.

그 결과는 MI 그래프에 대한 GL2vec-GRU 모형의 활용이 GL2vec 모형과 비교했을 때 성능 개선에 기여할 가능성이 있음을 확인하였다. 이는 종가 데이터의 그래프적 특성을 금융 시계열 예측에 있어서 중요한 정보를 제공한다는 것을 강조한다.

6. 결론

결론적으로, 본 연구는 종가 데이터를 바탕으로 한 종가 네트워크를 활용하는 GL2Vec(Graph convolutional network)을 이용한 ETF 가격 예측 모형을 제안하였다. 본 연구의 모형은 다양한 시장 요인 간의 복잡한 관계를 구조화 및 시각화함으로써 비전문가에게 금융 시장의 수익률 상호 정보량에 대한 이해도를 제고하였으며, 이를 통해 개선된 예측 결과를 도출하였다. 이러한 접근법은 금융 시장의 상황에 대한 이해도와 정확성을 높이는 데 큰 기여를 할 수 있다.

본 연구의 핵심 기여 중 하나는 금융 데이터를 GL2Vec 을 이용하여, 상호 정보량 지표를 바탕으로 복잡한 금융 상품과 시장 변수 간의 관계를 명확히 해석할 수 있도록 하는 것이다. 이는 금융 분석에 새로운 차원의 특성 또는 기술적 분석 접근 방법을 도입하는 것으로 볼 수 있다.

본 연구의 한계점은, 주로 종가 데이터만을 중심으로 한다는 점이다. 거시 경제 지표나 소셜 미디어, 뉴스의 감성 데이터와 같은 다양한 대체 데이터의 중요성은 이미 많이 알려져 있지만, 본 연구에서는 이를 고려하지 않았다. 또한, GL2Vec 의 사용에는 계산 시간 및 컴퓨팅 자원의 제약, 확장성에 대한 문제점이 있다. 이에 대한 보다 강건한 평가와 해결 방안 제시가 필요하다.

향후 연구 방향은, 제안된 모형을 다양한 금융 상품과 다른 국가의 시장에 적용하는 것이다. 이를 통해 모형의 일반화와 잠재적 적용 가능성을 검토하고자 한다. 추가로, 다양한 대체 데이터의 통합을 통해 예측력을 더욱 향상시키고, GL2Vec 을 중심으로 한 ETF 가격 예측 모형의 성능과 범용성을 높이려는 노력을 계속해나갈 것이다.

참고 문헌

Chen, H., & Koga, H. (2019). Gl2vec: Graph embedding enriched by line graphs with edge features. In Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26 (pp. 3-14). Springer International Publishing.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Hacine-Gharbi, A., & Ravier, P. (2021). On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion recognition. Journal of King Saud University-Computer and Information Sciences, 33(9), 1074-1083.

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