암호화폐 가격 정보 데이터에 대한 상관관계분석 및 회귀테스트

권도형*, 허주성**, 김주봉**, 임현교*, 한연희**†
*한국기술교육대학교 창의융합공학협동과정
**한국기술교육대학교 컴퓨터공학과
e-mail : {dohk, chil1207, rlawnqhd, glenn89, yhhan}@koreatech.ac.kr

Correlation Analysis and Regression Test on Cryptocurrency Price Data

Do-Hyung Kwon*, Joo-Seong Heo**, Ju-Bong Kim**, Hyun-Kyo Lim*, Youn-Hee Han**
Korea University of Technology and Education, Republic of Korea

요 약

기존의 전통적인 금융 시장에 대한 탐색적 데이터 분석에 비해 암호화폐에 대한 탐색적 데이터 분석은 전무하다. 본 논문에서는 대표적인 암호화폐인 비트코인을 비롯하여 총 12 개의 암호화폐에 대한 상관관계분석 및 회귀 모델을 적용하기 적합한지 여부를 결정하는 평균회귀테스트를 수행하고 그 결과에 대해 논한다.

1. 서론

최근 암호화폐들에 대한 가격을 예측하기 위해 다양한 접근들이 시도되고 있다. 특히 기존의 전통적인 주식시장에서의 주가 예측 기법들을 적용하려는 시도부터 딥러닝과 강화학습을 적용하려는 시도까지 다양하다 [1][2]. 현재 주식시장에서는 이미 사람이 직접 투자하는 방식이 아니라 알고리즘 트레이딩 프로그램을 활용하여 수익을 창출하는 것이 일반화 되어있다. 암호화폐 시장은 비트코인이 처음 등장한 이후의 짧은 역사를 갖고 있다. 그렇기 때문에 알고리즘 트레이딩 방식으로 암호화폐의 가격을 예측하기 위해서는 먼저 암호화폐 가격 데이터의 특성에 대한 고찰이 필요하다. 기존의 주가 분석과 비교할 때 암호화폐 시장에서 거래되는 암호화폐들에 대한 데이터 분석은 전무하다. 본질적으로 주가 데이터와 암호화폐 가격 데이터는 시계열 데이터로서 무작위적인 특성을 갖고 있으나, 암호화폐 가격 데이터가 더욱 변동성이 크며 특정 암호화폐의 가격에 의하여 다른 암호화폐의 가격이 영향을 받는다는 특징이 있다.

따라서 본 논문에서는 특정 기간에 대하여 수집된 암호화폐 가격 데이터에 대한 탐색적 데이터 분석(EDA)을 시도함으로써 암호화폐별 상관관계를 분석한다. 또한 기존의 주가 분석 기법으로 종종 쓰이는 회귀 모델을 적용하기 적합한지에 대한 평균회귀테스트를 수행한 결과를 살펴본다.

2. 데이터 수집 및 전처리

수집한 데이터는 빗썸 거래소 [6]가 제공하는 API 를 이용하였으며, { unix timestamp, 날짜 문자열(한국시), opening price, closing price, high price, low price, volume } 을 feature 로 갖는 10 분봉 데이터를 수집하였다. 수집의 대상이 된 암호화폐들은 빗썸 거래소에서 거래되는 12 개의 암호화폐인 BTC, ETH, XRP, BCH, LTC, EOS, DASH, XMR, ETC, QTUM, BTG, ZEC 이다.

<표 1> 수집된 암호화폐 가격 데이터 기간

코인	기간	데이터 개수
BTC	2017-06-09 08:50:00 ~ 2018-02-21 08:40:00	37008
ETH	2017-06-09 09:00:00 ~ 2018-02-21 08:40:00	37007
XRP	2017-06-09 09:00:00 ~ 2018-02-21 08:50:00	37008
BCH	2017-08-04 21:40:00 ~ 2018-02-21 08:40:00	28867
LTC	2017-06-09 09:00:00 ~ 2018-02-21 08:50:00	37008
EOS	2017-12-13 20:50:00 ~ 2018-02-21 08:50:00	10009
DASH	2017-06-09 09:00:00 ~ 2018-02-21 08:50:00	37008
XMR	2017-08-28 14:40:00 ~ 2018-02-21 08:50:00	25454
ETC	2017-06-09 09:00:00 ~ 2018-02-21 08:50:00	37008
QTUM	2017-10-20 18:10:00 ~ 2018-02-21 08:50:00	17801
BTG	2017-11-24 14:10:00 ~ 2018-02-21 08:50:00	12785
ZEC	2017-09-28 18:10:00 ~ 2018-02-21 08:50:00	20969

† 교신 저자: 한연희(한국기술교육대학교)

<표 2> 수집된 데이터 내부 모습의 예시

unix timestamp	날짜 문자열(한국시)	opening price	closing price	high price	low price	volume
1496326800000	2017.6.1 23:20	3094000	3118000	3127000	3094000	191.580257
1496327400000	2017.6.1 23:30	3116000	3119000	3124000	3116000	52.9553695
1496327940000	**2017.6.1 23:39**	3119000	3116000	3119000	3116000	63.8428308
1496328600000	2017.6.1 23:50	**0**	**0**	**0**	**0**	**0**
1496329854000	**2017.6.2 0:19**	**0**	**0**	**0**	**0**	**0**
1496331000000	2017.6.2 0:30	3100000	3080000	3124000	3078000	230.057336
1496331600000	2017.6.2 0:40	3081000	3090000	3100000	3081000	131.933399

<표 3> 이상치 데이터가 처리된 모습

unix timestamp	날짜 문자열(한국시)	opening price	closing price	high price	low price	volume	label
1496326800000	2017.6.1 23:20	3094000	3118000	3127000	3094000	191.580257	normal
1496327400000	2017.6.1 23:30	3116000	3119000	3124000	3116000	52.9553695	normal
1496328000000	**2017.6.1 23:40**	3119000	3116000	3119000	3116000	63.8428308	**adjustment**
1496328600000	2017.6.1 23:50	**3119000**	**3116000**	**3116000**	**3116000**	**3116000**	**zero**
1496329200000	2017.6.2 0:00	**3119000**	**3116000**	**3116000**	**3116000**	**3116000**	**(no data)**
1496329800000	2017.6.2 0:10	**3119000**	**3116000**	**3116000**	**3116000**	**3116000**	**(no data)**
1496329860000	**2017.6.2 0:20**	**3119000**	**3116000**	**3116000**	**3116000**	**3116000**	**adjustment+zero**
1496331000000	2017.6.2 0:30	3100000	3080000	3124000	3078000	230.057336	normal
1496331600000	2017.6.2 0:40	3081000	3090000	3100000	3081000	131.933399	normal

수집된 암호화폐들의 기간은 BCH, EOS, XMR, QTUM, BTG, ZEC 의 경우, 빗썸 거래소에 상장된 날짜에 따라 각각 다르며, 나머지 암호화폐들의 경우 시작된 날짜가 동일하게 2017년 6월 9일부터 시작되는데, 이는 빗썸 거래소 API 를 통해 데이터를 수집하는 시점에서부터 최대한의 과거이며, 이보다 더 과거의 데이터를 수집할 수는 없었다. <표 1>에 12 개의 암호화폐 각각의 기간을 나타내었다.

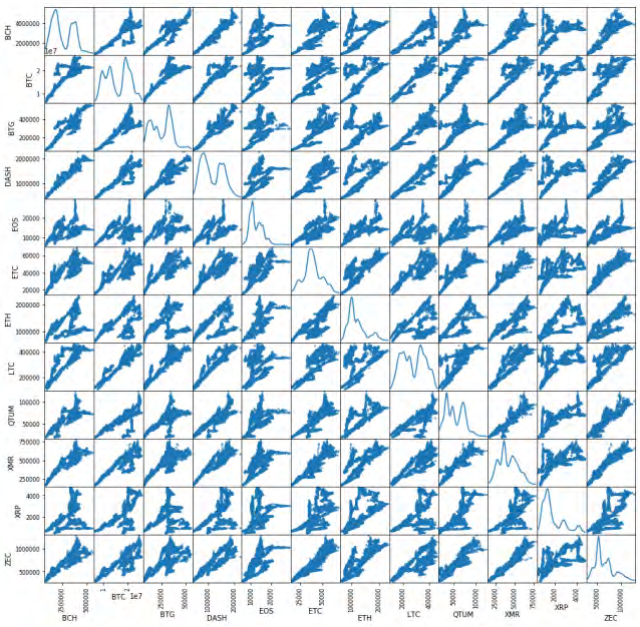
<표 2>는 수집된 데이터의 일부분을 예로 든 모습이다. 데이터의 내부에는 정상 데이터, 값이 0인 데이터, 600,000 unix timestamp 에 정확히 맞춰지지 않은 채 수집된 데이터가 보이며, 데이터 자체가 존재하지 않는 missing data 인 경우도 있다

<표 3>은 이러한 이상치 데이터들을 처리한 모습이다. 정상 데이터가 아닌, 언급한 세 가지 경우에 대하여 zero, adjustment, no data 로 라벨을 붙였으며, 정상 데이터에 대하여는 normal 로 라벨링을 하였다. 이상치 값은 직전 normal 행의 데이터를 네 개의 feature 각각에 대하여 복사한 값으로 채워 넣었으며, 추가적으로 'label' column 에 이를 표시했다. 추후 해당 데이터를 이용하여 인공신경망 학습에 쓰고자 할 경우에는 '날짜 문자열(한국시)' column 과 'label' column 은 제외한다.

3. 분석 및 평가

3.1. 상관관계분석

암호화폐 시장에서 거래되는 암호화폐들은 서로 다른 가치관을 가지고 있지만, 그 가격은 독립적으로 형성되지 않는다. 가장 잘 알려진 암호화폐인 비트코인은 현재까지도 지배적인 암호화폐로서, 다른 암호화폐들의 거래에 쓰이고 있을 뿐만 아니라 그 가격대 형성에 영향을 주고 있다. 이에 두 암호화폐 쌍의 상관관계를 분석함으로써 특정 암호화폐의 가격이 다른 암호화폐의 가격에 영향을 주는 정도를 살피는 것이 유의미하다고 판단하였다. (그림 1)은 이를 시각적으로 살펴보기 위하여 금융 분석에 쓰이는 대표적인 라이브러리인 pandas를 이용하여 두 암호화폐를 하나의 쌍으로 직교좌표계에 12 by 12 산점도 행렬(scatter plot) [4]로 나타낸 결과이다. 모든 암호화폐들이 우상향의 경향을 보이고 있음을 알 수 있다. 그러나 산점도 행렬 만을 살펴보는 것으로는 암호화폐별로 상관의 정도를 구체적으로 확인하기 힘들다. 따라서 상관계수를 함께 살펴보아야 한다.



(그림 1) 12 개 코인에 대한 산점도 행렬

<표 4> 12 개 코인에 대한 코인별 상관계수

	BCH	BTC	BTG	DASH	EOS	ETC	ETH	LTC	QTUM	XMR	XRP	ZEC
BCH	1.000000	0.869600	0.888080	**0.978363**	0.424021	0.785826	0.477490	0.883707	0.901557	0.933523	0.641242	0.884724
BTC	0.869600	1.000000	0.913137	0.928358	0.324508	0.784967	0.390082	**0.957097**	0.790524	0.880124	0.610747	0.809337
BTG	0.888080	0.913137	1.000000	0.910383	0.376878	0.743432	0.278752	**0.914425**	0.739909	0.841260	0.416226	0.744078
DASH	**0.978363**	0.928358	0.910383	1.000000	0.400848	0.803932	0.455574	0.930775	0.899156	0.943447	0.665373	0.887397
EOS	0.424021	0.324508	0.376878	0.400848	1.000000	0.608902	**0.765748**	0.271937	0.564107	0.591614	0.428413	0.562269
ETC	0.785826	0.784967	0.743432	0.803932	0.608902	1.000000	0.715497	0.809291	0.801597	0.900049	0.653307	**0.900402**
ETH	0.477490	0.390082	0.278752	0.455574	**0.765748**	0.715497	1.000000	0.319628	0.654514	0.678696	0.711443	0.741886
LTC	0.883707	**0.957097**	0.914425	0.930775	0.271937	0.809291	0.319628	1.000000	0.768149	0.863139	0.514836	0.810884
QTUM	0.901557	0.790524	0.739909	0.899156	0.564107	0.801597	0.654514	0.768149	1.000000	**0.921537**	0.787480	0.894817
XMR	0.933523	0.880124	0.841260	0.943447	0.591614	0.900049	0.678696	0.863139	0.921537	1.000000	0.740138	**0.952234**
XRP	0.641242	0.610747	0.416226	0.665373	0.428413	0.653307	0.711443	0.514836	**0.787480**	0.740138	1.000000	0.749549
ZEC	0.884724	0.809337	0.744078	0.887397	0.562269	0.900402	0.741866	0.810884	0.894817	**0.952234**	0.749549	1.000000

상관계수를 살필 때 각 암호화폐의 가격이 다르므로 0과 1 사이로 정규화한 값을 사용하며, 암호화폐별로 기간이 다르므로 모든 암호화폐의 기간을 동일하게 맞추기 위하여 데이터의 개수가 가장 적은 EOS를 기준으로 2017년 12월 13일 8시 50분부터 2018년 2월 21일 8시 50분까지의 기간을 공통된 기간으로 설정하였으며 그 결과는 <표 4>와 같다.

상관계수의 절대값에 따라 두 암호화폐의 상관 정도를 파악할 수 있는데, <표 4>에 의하면 모든 암호화폐가 최소한 절대값 0.3 이상의 상관계수를 보이고 있으므로 뚜렷한 상관관계를 갖는다고 파악할 수 있다. 0.7 이상의 상관계수를 보일 경우는 강한 상관관계를 갖는 경우이다. 예를 들어 BTC와 EOS, BTC와 ETH과의 상관계수는 BTC와 다른 암호화폐들과 비교했을 때 상대적으로 낮은 것으로 나타났지만 0.3 이상의 상관계수값을 가지므로 상관관계가 전혀 없는 것은 아니라고 할 수 있다. 암호화폐별로 상관계수가 가장 높은 곳에 별도의 처리를 하였다.

각 암호화폐별 상관계수의 평균을 구한 결과는 <표 5>와 같으며, 맨 위의 순서대로 상관계수의 평균값이 높은 정도를 나타낸다. BTC는 가장 지배적인 암호화폐인 것에 비해 해당 기간에는 중간 정도의 상관계수를 가지며, XMR이 가장 높은 상관계수를, EOS가 가장 낮은 상관계수를 갖는다. 따라서 EOS는 12개 암호화폐 중 가장 독립적으로 가격대가 형성된다고 할 수 있으며, XMR의 가격은 다른 암호화폐들의 가격과 유사하게 움직이는 것으로 판단할 수 있다.

<표 5> 상관계수의 평균

코인	상관계수(평균)	코인	상관계수(평균)
XMR	0.854	BTC	0.772
ZEC	0.828	LTC	0.754
DASH	0.817	BTG	0.731
QTUM	0.810	XRP	0.660
BCH	0.806	ETH	0.599
ETC	0.792	EOS	0.527

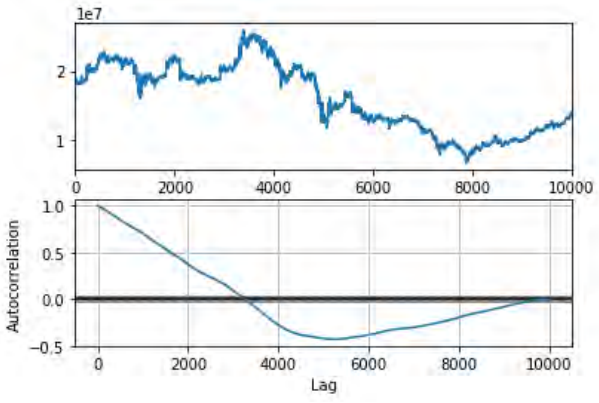
3.2. 자기상관관계

자기상관(autocorrelation)은 일정 시간 이후에 같은 패턴이 반복되는 것을 말한다. 어떤 데이터에 자기상관성이 보인다면 패턴이 있다고 할 수 있는데, 12 개의 암호화폐 모두 <표 1>에 해당하는 기간 동안에는 자기상관성이 없는 것으로 나왔다.

(그림 2)는 BTC 의 자기상관성을 살펴본 결과를 나타낸 그래프이다. BTC 이외의 다른 암호화폐들도 시간이 지남에 따라 자기상관성이 0 에 가까워지며, 모든 암호화폐들이 <표 1>의 기간 동안 (그림 2)에서 보는 바와 같이 자기상관성이 0 에 수렴하므로, 무작위성을 갖고 있다고 해석할 수 있다.

3.3. 평균회귀테스트

주가 데이터와 같은 시계열 데이터가 회귀 경향을 따른다는 단순한 가정을 세운다면 우리는 손쉽게 회귀 모델을 통해 주가 예측을 할 수 있을 것이다. 그러나 앞서 살펴본 바에 의하면 수집한 암호화폐 가격 데이터들은 무작위성을 띄고 있으며, 이는 평균으로 회귀하리라는 보장이 없음을 의미한다. 회귀 모델을 적용하기 위해서는 각 사건이 자신의 이전 사건에 영향을 받는, 비독립적인 사건이어야 한다. 따라서 회귀 모델을 적용하기 적합한지에 대한 여부를 결정하기 위하여 Augmented Dickey-Fuller 단위근 검정, Hurst Coefficient 분석, Regression half life 를 계산한 결과를



(그림 2) BTC 의 자기상관 그래프

간단히 살펴본다.

- Augmented Dickey-Fuller(ADF) 단위근 검정

ADF 단위근 검정은 Dickey-Fuller 검정에서 발전했다[5]. ADF 단위근 검정은 식 (1)의 모델로 표현되는 시계열 데이터에서 t시점의 데이터가 $t-1$시점의 데이터와는 독립적이라는 가설을 기각하는지를 테스트한다.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \varepsilon_t \qquad (1)$$

<표 1>의 기간에 대한 ADF 검정을 수행한 결과, BTC 의 경우 검정 통계량값이 -1.1268, 가설을 기각하기 위한 1% 기각값은 -3.4305, 5% 기각값은 -2.861 로서 가설을 기각하지 못하며, BTC 이외의 11 개의 암호화폐 모두에서 같은 결과를 보인다. 따라서 ADF 검정 결과에 의해 12 개의 암호화폐 모두 회귀 모델을 적용하기엔 적합하지 않다고 판단할 수 있다.

- Hurst Coefficient 분석

Hurst Coefficient 는 어떤 시계열 데이터의 확산 속도를 H 라는 표기로 나타냄으로써 무작위성으로 인해 예측 불가능한지, 시간이 지남에 따라 회귀하는 성질이 있는지, 아니면 높은 변동성을 띄는지를 판단할 수 있으며 그 정도를 나타낸다. <표 6>은 Hurst Coefficient 를 구한 결과이다. 12 개의 암호화폐 모두 0.5 보다는 작지만 0.5 에 가까운 값을 갖기 때문에 뚜렷한 회귀 성향을 보이지 않는다고 볼 수 있다.

- Regression half life

Regression half life 는 시계열 데이터의 평균으로의 회귀 주기를 찾는 방법이며 <표 1>의 기간에 대하여 구한 결과가 <표 6>에 포함 되어있다. 결과에 의하면 EOS 나 BTG 의 경우 회귀 주기가 상대적으로 짧기 때문에 회귀 모델을 적용하면 유의미한 결과를 얻을 수 있을 것으로 보인다. 반면 BTC, ETH, XRP 의 경우에는 상대적으로 그 값이 매우 높아서 회귀 모델은 적합하지 않은 것으로 보인다.

<표 6> Hurst Coefficient 분석 결과

코인	Hurst Coefficient	Regression half life
BTC	0.457698	**7857.5672**
ETH	0.441234	**5470.3199**
XRP	0.470470	**5388.8637**
BCH	0.465962	2990.2104
LTC	0.424898	5039.9336
EOS	0.444115	**349.3821**
DASH	0.428556	3759.4517
XMR	0.405608	2215.1896
ETC	0.443624	2369.9985
QTUM	**0.490686**	2218.8554
BTG	0.459099	**255.1854**
ZEC	**0.401158**	1437.1053

4. 결론

본 논문에서는 2017 년 6 월 9 일 8 시 50 분에서 2018 년 2 월 21 일 8 시 40 분까지의 기간(BTC 기준)에 대한 12 개의 암호화폐 데이터를 수집하고 전처리 과정을 거친 후, 해당 기간 동안의 암호화폐에 대한 탐색적 데이터 분석(EDA)과정을 통해 암호화폐끼리의 상관관계와 자기상관성을 분석하였으며, 회귀 모델에 적용하기 적합한 암호화폐를 찾기 위하여 ADF 단위근 검정, Hurst Coefficient, Regression half life 등을 살펴보았다. 그 결과 공통 기간인 2017 년 12 월 13 일 8 시 50 분부터 2018 년 2 월 21 일 8 시 50 분까지의 기간 동안 XRP 의 상관계수가 가장 높았으며 EOS 가 가장 독립적으로 움직였다. 또한 다른 암호화폐에 비해 회귀 모델을 적용하기 적합한 암호화폐로는 EOS 와 BTG 가 가능하다는 결론을 얻었다. 본 논문에서는 특정 기간을 한정하여 EDA 를 수행하였기 때문에 더욱 다양한 기간에 대한 분석을 수행할 필요가 있다. 추후 딥러닝 또는 강화학습 모델을 이용하여 암호화폐 가격 예측을 시도하고자 한다.

참고문헌

[1] Z. Jiang, D. Xu, and J. Liang, "A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem, " arXiv preprint arXiv:1706.10059, 2017.

[2] 송유정, 이재원, 이종우, "텐서플로우를 이용한 주가 예측에서 가격-기반 입력 피쳐의 예측 성능 평가, " 정보과학회 컴퓨팅의 실제 논문지, 제 23 권, 제 11 호, pp. 625-631, 2017.

[3] Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System, " 2008.

[4] S. B. Jarrell, Basic Statistics (Special pre-publication ed.), Wm. C. Brown Pub., 1994.

[5] D. A. Dickey and W. A. Fuller, "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," Journal of the American Statistical Association. Vol. 74, No. 366, pp. 427–431, 1979.

[6] https://www.bithumb.com/