|  |
| --- |
| 가상화폐에 대한 탐색적 데이터 분석 |
|  |
| 권도형\*, 한연희\*\*  \*한국기술교육대학교 창의융합공학협동과정  \*\*한국기술교육대학교 컴퓨터공학과  e-mail : {dohk, yhhan}@koreatech.ac.kr |
|  |
| Explonatory Data Analysis on Cryptocurrency |
|  |
| Do-Hyung Kwon\*, Youn-Hee Han\*\*  Korea University of Technology and Education, Republic of Korea |
|  |
| **요 약**  기존의 전통적인 금융 시장에 대한 탐색적 데이터 분석에 비해 가상화폐에 대한 탐색적 데이터 분석은 전무하다. 본 논문에서는 대표적인 가상화폐인 비트코인을 비롯하여 총 12개의 코인에 대한 상관관계분석 및 평균 회귀 모델을 적용하기 적합한지 여부를 결정하는 평균회귀테스트를 수행하고 그 결과에 대해 논한다. |
|  |

1. 서론

최근 가상화폐 중 대표적인 코인인 비트코인을 비롯한 가상화폐들에 대한 가격을 예측하기 위해 다양한 접근들이 시도되고 있다. 특히 기존의 전통적인 주식시장에서의 주가 예측 기법들을 적용하려는 시도부터 딥러닝과 강화학습을 적용하려는 시도까지 다양하다. 본 논문에서는 특정 기간에 대하여 수집된 코인 가격 데이터에 대한 탐색적 데이터 분석(EDA)를 시도함으로써 코인별 상관관계를 분석하고, 기존의 주가분석 기법으로 주로 쓰이는 평균회귀모델을 적용하기 적합한지에 대한 테스트를 수행한 결과를 살펴본다.

1. 관련연구

현재 주식시장에서는 이미 사람이 직접 투자하는 방식이 아니라 알고리즘 트레이딩 프로그램을 활용하여 수익을 창출하는 것이 일반화 되어있다. 최근에는 인공지능 기술의 발전에 의해 인공지능을 이용하여 주식시장에서 주가를 예측하고자 하는 시도들이 늘어나고 있다 [1][2]. 그러나 가상화폐 시장은 비트코인이 처음 세상에 나오기 시작한 이후의 짧은 역사를 갖고 있으며, 알고리즘 트레이딩 방식으로 가상화폐의 가격을 예측하기에 앞서 그 데이터의 특성에 대한 고찰이 먼저 필요하다. 가상화폐는 탈중앙화를 표방하는, 각국 중앙은행이 발행하는 화폐의 대안으로서, 현재까지 가장 잘 알려진 가상화폐로는 비트코인이 있다 [3]. 주가의 경향을 분석할 때는 보통 하나의 주가 종목에 대한opening price, closing price, high price, low price 를 이용하여 산점도를 그리고 상관계수를 살핀다. 이 때 상관계수의 부호는 두 변수가 함께 증가하거나 함께 감소하는 경향이 있을 경우 양수로 나타나게 된다. 또한 두 변수끼리 짝지어 그 상관관계를 직교좌표계에 산점도 행렬(scatter plot) [4]로 나타냄으로써 그 경향을 시각적으로 파악할 수 있다.

1. 데이터 수집

수집한 데이터는 빗썸 거래소가 제공하는 API를 이용하였으며, { unix timestamp, opening price, closing price, high price, low price, volume }을 feature로 갖는 10분봉 데이터를 수집하였다. 수집의 대상이 된 코인들은 빗썸 거래소에서 거래되는 12개의 코인인 BTC, ETH, XRP, BCH, LTC, EOS, DASH, XMR, ETC, QTUM, BTG, ZEC이다. 수집된 코인들의 기간은 BCH, EOS, XMR, QTUM, BTG, ZEC의 경우, 빗썸 거래소에 상장된 날짜에 따라 각각 다르며, 나머지 코인들의 경우, 시작된 날짜가 동일하게 2017년 6월 9일부터 시작되는데, 이는 빗썸 거래소 API를 통해 데이터를 수집하는 시점에서부터 최대한의 과거이며, 이보다 더 과거의 데이터는 수집할 수는 없었다. <표1>은 12개의 코인들 각각의 기간을 나타내고 있다.

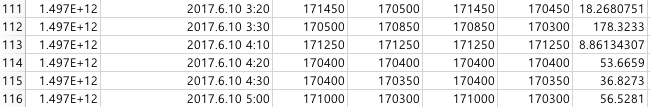
<표 1> 수집된 코인데이터 기간



(그림1)에 보이는 바와 같이 수집된 데이터의 내부에는 정상 데이터, 값이 0인 데이터들, 600,000 unix timestamp에 정확히 맞춰지지 않은 채 수집된 데이터들이 보이며, 추가적으로 (그림2)에 보이는 바와 같이 데이터 자체가 존재하지 않는 missing data인 경우가 있다. 정상 데이터 이외에 언급한 세 가지 경우에 대하여 zero, adjustment, no data로 라벨을 붙였으며, 정상 데이터에 대하여는 normal로 라벨링을 하였다. 추후 해당 데이터를 이용하여 인공신경망 학습에 쓰일 경우 라벨 column은 제외한다.



(그림 1) 수집된 데이터 내부 모습



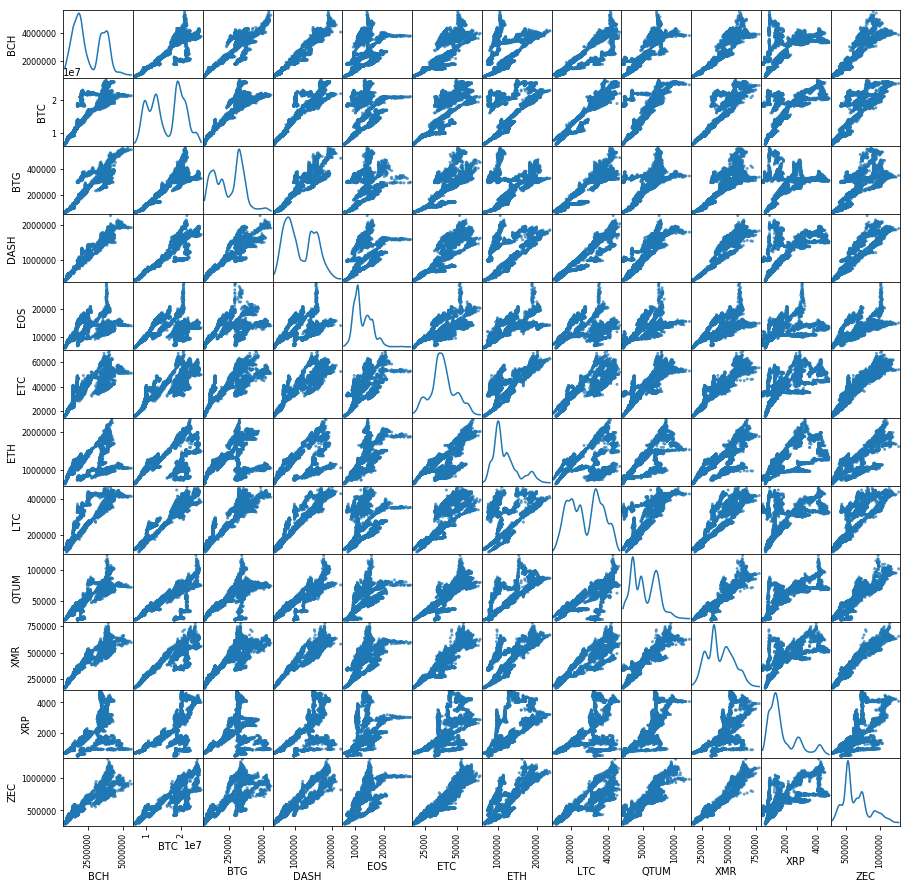
(그림 2) no data 상황의 데이터 모습

1. 분석 및 평가
2. 상관관계분석

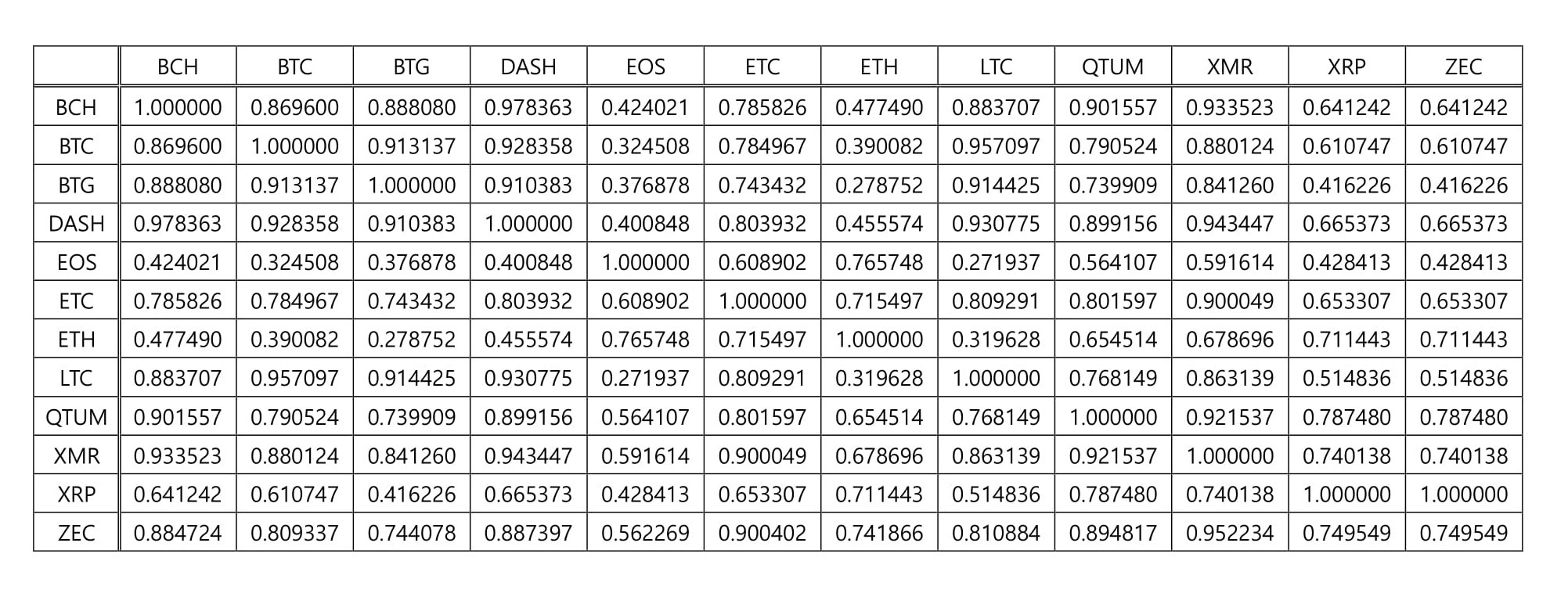
가상화폐 시장에서 거래되는 코인들은 서로 다른 가치관을 가진 코인들이지만 그 가격은 독립적으로 형성되지 않는다. 가장 잘 알려진 가상화폐인 비트코인은 현재까지도 지배적인 가상화폐로서, 다른 코인들의 거래에 쓰이고 있을 뿐만 아니라 다른 코인들의 가격대 형성에 영향을 주고 있다. 두 코인간의 상관관계를 직교좌표계에 산점도 행렬(scatter plot)로 나타냄으로써 그 경향을 시각적으로 파악할 수 있다. 금융 분석 및 데이터 처리에 쓰이는 대표적인 라이브러리인 pandas를 이용하여 산점도 행렬과 상관계수를 살펴본 결과는 (그림 3)과 <표 2>와 같다. 그러나 산점도 행렬 만을 살펴보는 것으로는 코인별로 정확한 상관관계를 확인하기 힘들며, 따라서 상관계수를 살핀다.

상관계수를 살필 때 각 코인의 가격이 다르므로 0과 1 사이로 정규화한 값을 사용하며, 코인별로 기간이 다르므로 모든 코인의 기간을 동일하게 맞추기 위하여 데이터의 개수가 가장 적은 EOS를 기준으로  2017년 12월 13일 수요일 오후 8:50:00 부터 2018년 2월 21일 수요일 오전 8:50:00까지의 기간을 설정하였다.

산점도 행렬을 통해 모든 코인들이 우상향의 경향을 보이고 있음을 알 수 있다. 상관계수의 절대값에 따라 두 코인의 상관정도를 파악할 수 있는데, <표 2>에 의하면 모든 코인이 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 코인들과 비교했을 때 상대적으로 낮은 것으로 나타났지만 0.3 이상의 상관계수값을 가지므로 상관관계가 아주 없는 것은 아니라고 할 수 있다.



(그림 3) 12개 코인에 대한 산점도 행렬

<표 2> 12개 코인에 대한 코인별 상관계수

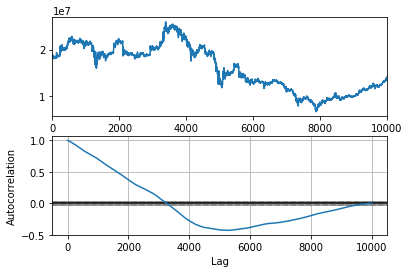
<표 3> 상관계수의 평균



각 코인별 상관계수의 평균을 구한 결과는 <표 3>과 같으며, 맨 위의 순서대로 상관계수의 평균값이 높은 정도를 나타낸다. BTC는 가장 지배적인 코인인 것에 비해 해당 기간에는 중간 정도의 상관계수를 가지며, XMR이 가장 높은 상관계수를, EOS가 가장 낮은 상관계수를 갖는다. 따라서 EOS는 12개 코인 중 가장 독립적으로 가격대가 형성된다고 할 수 있으며, XMR의 경우 XMR의 가격에 따라 다른 코인들의 가격이 따라 움직이는 것으로 판단할 수 있다.

1. 자기상관관계

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



(그림 4) BTC의 자기상관 그래프

(그림 4)는 BTC의 자기상관성을 살펴본 결과 그래프이다. BTC외의 다른 코인들도 시간이 지남에 따라 자기상관성이 0에 가까워지며, 모든 코인들이 <표 1>의 기간 동안 무작위성을 갖고 있다고 해석할 수 있다.

1. 평균회귀테스트

주가 데이터와 같은 시계열 데이터가 평균회귀 경향을 따른다는 단순한 가정을 세운다면 우리는 손쉽게 회귀 모델을 통해 주가 예측을 할 수 있을 것이다. 그러나 앞서 살펴본 바에 의하면 수집한 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 적용한지 적합한지에 대한 여부를 결정하기 위하여 Augmented Dickey-Fuller 단위근 검정, Hurst Exponent, regression half life를 계산한 결과를 살펴보고자 한다.

* Augmented Dickey-Fuller(ADF) 단위근 검정

ADF 테스트는 Dickey-Fuller 검정에서 발전했다 [5]. 식 (1)의 모델로 표현되는 시계열 데이터에서 시점의 데이터가 시점의 데이터와는 상관관계가 없다는 가설을 기각하는지를 테스트한다.

(1)

<표 1>의 기간에 대한 ADF 검정을 수행한 결과, BTC의 경우 검정 통계량값이 -1.1268, 가설을 기각하기 위한 1% 기각값이 -3.4305, 5% 기각값이 -2.8616으로서 가설을 기각하지 못하며, BTC 이외의 11개 코인 모두에서 같은 경향을 발견하였다. 따라서 ADF 검정 결과에 의해 12개 코인 모두 회귀 모델을 적용하기엔 적합하지 않다고 판단할 수 있다.

* Hurst Exponent
* regression half life

1. 결론

* **논문작성은 A4용지 2~4페이지까지만 허용 합니다.**

**(4페이지 초과시 게재되지 않습니다.)**

**참고문헌**

[1] G. Armano, M. Marchesi, A. Murru, A hybrid genetic-neural architecture for stock indexes forecasting, “ Information Sciences 170(1), pp.3-33, 2005.

[2] J. W. Lee, "고변동 주가 패턴의 감독 학습에 기반한 주식 거래 시스템, “ 정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 19 권 제 1 호(2013.1) p.23~29

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

[4] Jarrell, Stephen B. (1994). Basic Statistics (Special pre-publication ed.). Dubuque, Iowa: Wm. C. Brown Pub. p. 492

[5] Dickey, D. A.; Fuller, W. A. (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root". Journal of the American Statistical Association. 74 (366): 427–431