Предлог пројекта из СИАП-а

Овај документ садржи кратак опис онога што је тема пројекта и дефиниција, мотивација за одабрану тему. Након мотивације следи преглед литературе, затим скуп података који је укратко описан. Такође је наведен и софтвер који ће бити коришћен, као и метод евалуације.

Дефиниција пројекта (циљ)

Криптовалуте су због скорашњег ценовног краха у жижи јавности. Сходно томе, велики број научних радова и истраживања фокусира се на утврђивање одређених правила флуктуације њихове цене.

Тема пројекта је поређење пројекције вредности криптовалута Etherium и Bitcoin ради лакшег одабира потенцијалног улагања у исте.

Мотивација

Криптовалуте као модерно и високоризично тржиште заокупљују пажњу великог броја инвеститора који због саме природе наведеног тржишта, немају искуства и потребно знање за адекватно улагање. Иако постоји велика количина софтвера која се бави сличним проблемима, њихово поље је сама предикција индивидуалних вредности криптовалута, док се решење овог пројекта бави поређењем вредности и предлогом повољнијих опција улагања.

Литература

1. *S., Kim, J., & Im, H. (2019). A comparative study of bitcoin price prediction using deep learning. Mathematics, 7(10), 898*

   Тема рада: Предикција вредности криптовалуте биткоин коришћењем Deep Learning алгоритама, као и поређење добијених резултата

   Подаци: Преузети са сајтова Bitstamp Bitcoin market time series data и Bitcoin blockchain information and its various statistics. Одабрано 18 релевантних атрибута чији је Spearman корелациони фактор у интервалу [0.75, 0.95]. Вредности наведених атрибута прикупљане су на дневном нивоу у временском интервалу од 2590 дана. При процесирању дате вредности су нормализоване и издељене у секвенце при чему је 80% података коришћено за обучавање док је преосталих 20% коришћено за тестирање.

   Алгоритми:

   - DNN
     - Улаз у мрежу представљају претпроцесиране вредности претходно поменутих 18 атрибута у периоду од m дана док излаз из исте представља вредност биткоина у m+1 дану. У случају класификационог проблема добијена вредност биткоина трансформисана је коришћењем сигмоидне функције и заокружена на конкретну вредност 1 (вредност расте) односно 0 (вредност опада). Коришћена је активациона функција ReLU, а за валидацију резултата коришћена је метода средње квадратне грешке.

- RNN i LSTM (Long Short-Term Memory)
    - Улази и излази из мреже су исти као и у претходном случају, разлика између DNN i RNN је у томе што RNN уводи стање у систем те при транзицији из стања у стање осим стандардних функција и коефицијената утичу и претходна стања. Надоградња над RNN представљају LSTM које решавају проблем „краткорочног“ памћења стања односно да што је даље претходно стање то мање утиче на наредно и при већем броју улазних података њихова међузависност се губи (проблем нестајућег градијента). LSTM уводи „гејтове“ улаза, излаза и заборава као и вектор стања чиме одржава међузависност стања при великом броју улаза.
- CNN
    - Улази и излази су исти као и у претходним случајевима. Коришћена је једноставна конволуциона мрежа са једним 2D конволуционим слојем са 36 филтера величине 3 x 18 (број улаза). Број филтера добијен је експерименталном методом при чему је такође утврђено да додавање додатних слојева не побољшава резултате.
- DRN
    - Надоградња над претходно поменуту CNN тиме што решава претходно поменути проблем нестајућег градијента увођењем резидуалних блокова. Архитектура коришћене мреже састоји се од улазног слоја, једног конволуционог 2D слоја, 4 резидаулна блока, једног “pooling” слоја и излазног потпуно-повезаног слоја.

Резултати: Класификациони модели, који дају вредности 1 (пораст вредности) односно 0 (пад вредности) показали су се ефективнијим за алгоритамско трејдовање. DNN алгоритам је резултовао најмањом средном вредошћу квадрата грешке при решавању класификационог проблема, а LSTM при решавању регресивног проблема.

2. *Kang, C. Y., Lee, C. P., & Lim, K. M. (2022). Cryptocurrency Price Prediction with Convolutional Neural Network and Stacked Gated Recurrent Unit. Data, 7(11), 149.*

Tema rada: Предлагање хибридног модела за предикцију цене криптовалута који се састоје од " 1-dimensional convolutional neural network" и " stacked gated recurrent unit", заједно назван скраћеницом 1DCNN-GRU. Идеја је да 1DCNN даје високо дискриминативне податке, док GRU хвата далекосежне зависности у подацима.

Podaci: Коришћени су историјски подаци три криптовалуте и то Bitcoin, Etherium и Ripple. Након добављања подаци су претпроцесирани како би се надоместили недостаци. Подаци о биткоину су добављени са Kaggle сајта и представљају вредности снимљене са периодом од 1 минута од 1. Јануара 2012 до 31. Марта 2021. Састоје се од око 4.8 милиона узорака међу којима има NaN вредности. Неке од података које сваки узорак има су "open", "high", "low", "close (OHLC)", "volume" и "weighted price" као и време сваког читања који је у UNIX формату. Етереиум подаци су добављени са Bitstamp exchange сајта. Састоје се од око

396403 узорака са интервалом од 1 минут. Рипл подаци су такође добављени са Bitstamp exchange сајта и има исти број узорака као и у случају етериума са истим интервалом. Што се претпроцесирања тиче идеја је била да се обраде историјски подаци како би се спремили за неуронску мрежу. Претпроцесирање се састојало од избора одлика које су од значаја, конверзије времена из UNIX у YY:MM:DD, уклањање вредности које недостају, раздвајања на обучавајући и валидациони скуп и мин-мах скалирања ради нормализације.

Algoritmi:

- 1DCNN (1-dimensional convolutional neural network) i GRU (stacked gated recurrent unit). Мрежа је постављена тако да има један слој 1DCNN и два GRU слоја која имају по 256 " units". Употреба чистих података о ценама би унела велику количину шума и одступања чиме би изазвала сметње при учењу самог регресионог модела. 1DCNN ту стоји како би извукао правило у историјским подацима цена. У овом слоју кернел клизи по временској оси и представља подакте по њиховим одликама. Након тога два GRU слоја треба да извуку дугорочна правила на основу извучених одлика. То је могуће због две "капије", "update" капија и "reset" капија. На самом крају излаз из GRU слоја се пушта у "dense" слој са једним скривеним "unit"-ом за предикцију цене. Хиперпараметри који су подешавани за време алгоритма су: оптимајзер, активациона функција и бечинг. Поређена су четири оптимајзера, и то Adam, SGD, Adamax и RMSProp. Од активационих функција разматране су сигмоид, softmax, ReLU, tanh и линеарна као и различите стратегије бечинга. Најбољи резултат показала је комбинација SGD-а, сигмоидне активационе функције и величине беча од 16

3. *Zhang, S., Li, M., & Yan, C. (2022). The Empirical Analysis of Bitcoin Price Prediction Based on Deep Learning Integration Method. Computational Intelligence and Neuroscience, 2022.*

Тема рада: Идеја је да се користе две комбиноване технике за предикцију цена криптовалута. Ове две технике су напредна "deep neural network model" (stacking denoising autoencoders) SDAE а друга је bootstrap aggregation (Bagging).

Подаци: Коришћени су разни подаци везани за биткоин. Неки од тих податак су величина ланца, хеш рејт, тежина мајнинга, број трансакција и вредност и ти подаци су преузети са сајтова Data.Bitcoin.org, Blockchain.com и CoinMarketCap. Поред тога Baidu и Google претраге везане за биткоин преузете су са Baidu Index-а и Google trends-а. Одређени релевантни догађаји су такође узети у обзир. Они су прикупљени ревизијом 11-те годишњице биткоина од стране новина " Daily star". Цена злата је преузета са Goldhub -а а вредност долара је преузета са Investing.com-а. Сви подаци су опсегу од 29. Новембра 2014 до 30. Марта 2020. Ово истраживање препознаје 9 фактора биткоина као одговорно за његову цену и то су величина ланца, хаш рејт, тежина мајнинга, волумен размене, вредност на маркету, Baidu и Google претраге везане за биткоин, цену злата, индекс долара и узима у обзир велике догађаје везане за крипто свет. Разлози за избор баш ових

фактора су ти да је веза између цене биткоина и ових фактора врло нелинеарна и пуна флуктуација али свака појединачно може да донесе битне информације о цени биткоина.

Алгоритми: Stacking denoising autoencoders је популарни DNN модел и резултати показују да има већу тачност при процени од стандардних модела машинског учења. То постиже тако што додаје неколико аутоенкодера који уклањају шум. Сам аутоенкодер јесте једнослојна неуронска мрежа са истим бројем улаза и излаза. Denoising аутоенкодер додаје шум на податке како би се смањио overfitting проблем што побољшава робустност. Додавањем више оваквих аутоенкодера добија се "stacked denoising autoencoders" архитектура. Сваки слој врши ненадгледано учење одвојено како би се смањила грешка између улаза и резултата. Тек када се к-ти слој обучи к+1 слој може да се обучава јер се користи пропагација унапред те је излаз из К-тог слоја улаз у К+1 слој.
Bagging је скраћеница за bootstrap aggregating и користи се за проблеме класификације и регресије. Ако претпоставимо да има м узорака, тада бирамо један узорак као улаз и онда га враћамо назад у скуп свих узорака како бисмо очували вероватноћу да за наредно узорковање. Коришћен је и модел мултиваријантног предвиђања. За разлику од time-series модела овај модел разматра и ауторегресивне ефекте саме серије података као и утицај егзогених променљивих. Комбиновани алгоритам се састоји од наредних корака:

- Претпроцесирање - трансформација оригиналних података у тренинг и тест скуп
- Мултипликација тренинг скупа - генериши К самплова уз помоћ Bootstraping алгоритма
- Тренирање - свака група самплова тренира К SDAE модела
- Сумирање резултата - узми средњу вредност К вредности као коначни резултат

Резултати: Како би се евалуирале перформансе користе се три најчешће коришћена начина и то:

- Direction accuracy (DA)
- Mean absolute error (MAPE)
- Mean square root error (RMSE)

У поређењу са LSSVM и BP као традиционалним моделима учења, SDAE-B метод рамзатран у овом раду се показао као бољи са већом прецизношћу и мањом грешком.
MAPE износи 0.016, RMSE износи 131.643 и DA износи 0.817

Скуп података

Због саме природе крипта, односно његове динамичне свакодневне промене, неопходна је велика количина података како би се могло радити процењивање вредности. Ти подаци ће бити преузети са coinmetrics сајта. Биткоин и етериум због своје популарност обилују подацима који се мере. Подаци о биткоину су разматрани у распону од 3. Јануара 2009. до 9. Јануара 2023. са интервалом од један дан, док су подаци о етериуму узети у распону од 30. Јула 2015. до 9. Јануара 2023. такође са интервалом од један дан. Изобиље могућих информација везаних за посебне валуте навеле су на избор следећих одлика:

1. Time: Датум када је читање направљено. Како ће модел бити "time-series" ово ће нам бити репер за timestamp
2. BlkCnt: представља број блокова направљених на дан читања
3. CapMrktCurUSD: представља суму тренутног стања у доларима
4. DiffMean: средња вредност тежине проналаска хеша блока
5. FeeMeanUSD: средња вредност трошка по трансакцији у доларима
6. FlowInExUSD: укупно размењене криптовалуте међу корисницима у доларима
7. HashRate: средње стопа решавања хешева
8. NDF: Network distribution factor, однос стања који држе адресе са макар 1/1000 од текуће јединице валуте
9. ROI30d: повратна вредност за једну јединицу валуте ако је купљена у претходних 30 дана

Због велике разлике у значењу а и вредностима података претпроцесирање ће бити неопходно. Подаци као што су укупна количина размењених јединица валуте и средња вредност тежине проналаска хеша блока су подаци који имају велику флуктуацију на почетку и на крају сета те је неопходно применити неки од метода за нормализацију. Неки атрибути имају недостајуће вредности за неке податке на почетку сета као нпр ROI30d и FlowInExUSD те ће они бити замењени са 0. Поред поменутих током рада биће уочене и документоване све додатне методе претпроцесирања.

Сет ће бити подељен на тренинг сет и сет за тестирање. Како ће се обучити две мреже, једна за биткоин друга за етериум за обе ће, од својих респективних сетова, бити издвојено 70% за сет за обучавање и 30% за сет за тестирање. Излаз из система ће бити представљен као једна вредност - PriceUSD (цена у доларима на крају посматраног дана).

При обучавању модела примениће се надгледано учење, односно моделу ће за одређене улазе бити прослеђени и кореспондентни излази. Идеја је да се налик у 1. раду улазни подаци сведу на секвенцу од m дана (при чему је улаз m x 9 (број релевантних атрибута)) док се на излазу посматра вредност биткоина у m+1 дану. Временски интервал m биће утврђен емпиријски, при имплементацији и обучавању-тестирању модела.

Алгоритми, методологија и евалуација

Идеја је да се тестира комбиновани алгоритам из другог рада 1DCNN-GRU. У овом раду аутор је користио уносе који су везани стриктно за вредност криптовалута док ће се у овом раду размотрити проширење скупа одлика (feature-а) који би могли утицати на цену што је преузета идеја из трећег рада. Сама мрежа ће изгледати исто:

1. 1DCNN (1-dimensional convolutional neural network) - један слој, 256 филтера, величина кернела 1, корак 1
2. GRU (stacked gated recurrent unit) - два слоја по 256 чворова
3. Густи слој - 1 чвор

Што се хиперпараметера тиче, као и поменутом раду са најбољим резултатом, биће коришћена сигмоид активациона функција, величина беча 16 и SGD оптимајзер за мрежу која се обучава за процену вредности биткоина, док ће комбинација Adamax оптимајзера, softmax активационе функције и величине беча 32 бити коришћена за етериум. Као методе валидације резултата биће коришћене RMSE и R2.

Софтвер

Пројекат ће бити имплементиран коришћењем програмској језика Python и библиотека за Deep learning које он нуди.

Тим

Алексић Никола E2-99/2022, Милосављевић Никола E2-36/2022, Врбица Владо E2-95/2022