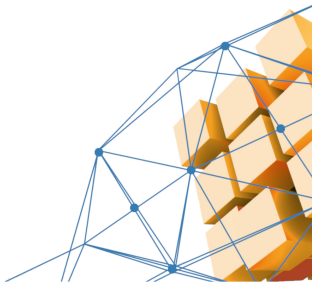
The links between Machine Learning and Blockchain

Andrea Gangemi, 15 May 2020



Motivation

- *Machine Learning* (ML) and *Blockchains* are two of the most disruptive technologies of the last years.
- A blockchain can be seen like a decentralized database, while Machine Learning is useful to find special patterns on data.
- For this reason, it is possible to integrate them together: however, until now there has been little work towards this objective.
- We describe what has been done during these years to put together these two technologies.

What is Machine Learning?

- Machine Learning is a method of training algorithms which improve through experience, mainly used for understanding data.
- ML algorithms build a statistical model based on known data, in order to make predictions on future, unseen data.
- Usually, ML algorithms are divided into *regression* algorithms and *classification* algorithms.
- Machine Learning should not be confused with Artificial Intelligence or Deep Learning.

Supervised vs Unsupervised

ML algorithms can also be classified based on how they act on data.
They can be divided in:

- *Supervised Learning*: it builds a mathematical model on a set of data which contains informations about the inputs and the desired output.
  The dataset used to train the model is divided into the *train set* and the *test set*.

- *Unsupervised Learning*: it takes a set of data which contains only inputs (usually non labeled) and try to find some structure on these data.

The Standard Approach

A typical ML approach consists in:

- Analyze the dataset with unsupervised techniques;
- Use supervised algorithms to predict a label or a variable.

Given its properties, a blockchain can be an instrument to improve machine learning: for example, it can be used to store datasets.

Applications

There are four main applications which connect ML with Blockchains:

- *Classification* of Blockchain users.
- *Prediction* of cryptocurrencies prices.
- *Saving* of ML datasets exploiting smart contracts features.
- A *Consensus Algorithm* based on ML.

Let's describe the main results obtained during the last years towards these different research fields.

Address Security

- On a blockchain, the user identity is protected by an address.

However, a careful analysis gives the possibility to recover most of the addresses used by a specific user.

- Some addresses are publicly written online: exchanges or gambling sites.
- Some people used to write their Bitcoin address on forums to advertise cryptocurrencies.
- Some people published accidentally their address asking for help in technical forums.

Address Clustering

Exploiting the structure of a transaction, it is possible to classify every user on the blockchain, starting from known addresses.

The main hyphothesis under this approach are:

- Every user belongs to an unique group.
- The addresses used as inputs in a transaction all belong to the same user.
- If there are more than two addresses in the output of the transaction, one of them belongs to the user who started it.

Address Clustering

Thanks to *clustering techniques*, it is possible to recover the set of addresses belonging to a particular user.

The main approach is based on transaction inputs.

- With this ML approach, it is possible to get big groups of addresses which are all linked to the same user.
- Analyzing the outputs of their transactions, we can classify some of these groups.
- This does NOT mean we can go back to the real identity of an user behind these addresses.

Supervised Learning

After the clustering, supervised learning algorithms can be used to label the groups which are still not classified, starting from the known clusters.

- Group labels are taken from a specific set: around half of them are linked to illicit activities.
- The known clusters are divided into two sets, the *training set* and the *test set*.
- Every cluster also contains informations about the transactions sent by every address.

Cluster Categories
darknet-market
exchange
gambling
hosted-wallet
merchant-services
mining-pool
mixing
other
personal-wallet
ransomware
scam
stolen-bitcoins

Supervised Learning

Several supervised algorithms can be used on the prepared dataset, and their performance can be computed, for example, with the *F1-score*.

- Every algorithm learns from data in the training set, and its accuracy is then proved on the test set.
- According to the majority of the studies, the best performing algorithm is the *Gradient Boosting Classifier*.

Limitations of the Approach

- Some classes are unbalanced, meaning that it is difficult to build an accurate predictor for every class. Oversampling techniques could be applied.
- The number of categories could not be enough to represent the whole network.
- Even though it is a good starting point, having each cluster belonging to an unique label is a strong limitation.

Price Prediction

Another interesting ML application on blockchains is the forecast of future cryptocurrencies prices.

Hypothesis in most price prediction models are:

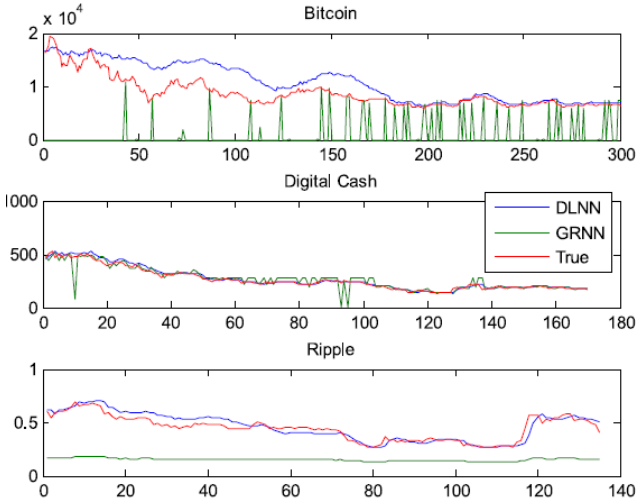
- Crypto prices are taken by a single exchange.
- The daily price is an average of the daily price fluctuations.

In this case, the performance metric is the *RMSE*.

Price Prediction

- A lot of supervised algorithms have been tested; the best results have been produced by *Gradient Boosting* and *Neural Networks*.
- Gradient Boosting is more efficient for short-term predictions, while Neural Networks are more efficient with bigger time intervals.
- Neural networks works better with millions of data, so it is expected to have better price predictions in the next years, for both long and short-term predictions.

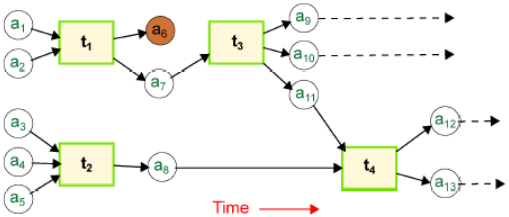
Price Prediction



Price Prediction with Graphs

An interesting idea to improve price predictions exploits the properties of the *graph structure* of a blockchain.

- In graph terminology, addresses and transactions are nodes, while a transfer of coin represents a link.



Price Prediction with Graphs

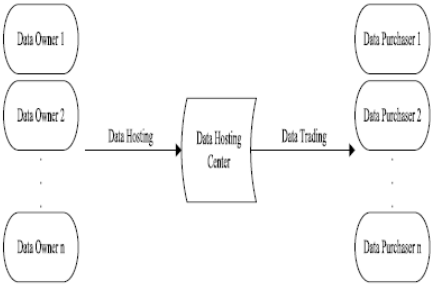
The study focuses on particular subgraphs, composed by a single transaction and any number of input and output addresses.

- Subgraphs are clustered using *Cosine Similarity*.
- Various ML models are tested on both the whole dataset and the chosen subgraphs.

The ML models computed on the chosen subgraphs had more accurate price predictions, especially for long-term forecasts.

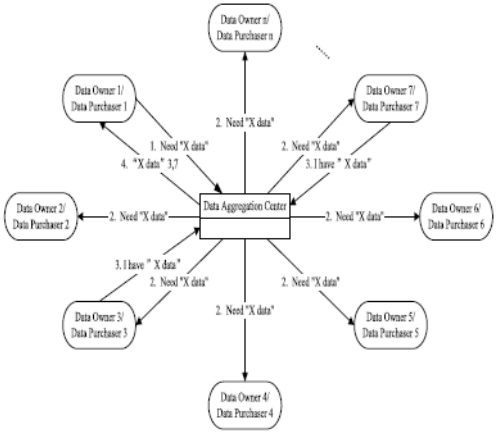
Decentralized ML

- In the big data era, data are usually collected and stored into one central server.
- People who want to process data contact the center to acquire them.
- This model is not safe because the trading center completely owns the data.



Decentralized ML

- A first solution is the management of the data by the owners themselves.
- The central server now has only the job to connect data owners and data purchasers.
- This model is still not safe, because the trading center has the ability and the opportunity to retain the trading data.



Decentralized ML

A blockchain can achieve the decentralization to improve the traditional data trading modes.

- Two characteristics of the blockchain are *immutability* and *tamper-proof*.
- Some blockchains also allow the use of *smart contracts*.

A possible solution exploits these key properties to build a decentralized framework.

However, blockchains are not enough because datasets are too heavy to be saved on them.

Decentralized ML

For this reason, the proposed solution utilizes also the InterPlanetary File System (IPFS), which stores data in a distributed way.

Suppose now every actor has an address on a smart contract blockchain, e.g. Ethereum. Then:

- Data are saved on the IPFS server.
- The hash of data, provided by the IPFS itself, is saved into the blockchain.
- The data center creates a smart contract, which, among other things, records the owner and the hash of a dataset.

Decentralized ML - Steps

- The data purchaser sends a request to buy some data off-chain.
- A data owner with the required data contacts the purchaser and tells him its address.
- The purchaser checks if the address exists on the contract.
- The purchaser, through the contract, asks the owner to send some specific data.
- The owner and the purchaser sends the same amount of Ether to the contract.

Decentralized ML - Steps

- The contract generates a token which gives to the purchaser the right to download the data.
- After the successful download, the purchaser confirms the end of the trade.
- The deposit of the owner is given back, while the deposit of the purchaser is distributed between the owner and the center.

What happens if the dataset is not what the purchaser was expecting?

Proof-of-Learning

- Machine learning models exhibit a property that is desirable for a Proof-of-Work consensus mechanism: they are hard to solve but easy to verify.
- *Proof-of-Learning* (PoL) takes inspiration from Kaggle competitions, which allow the development of new performing algorithms in relatively small periods of time.
- The most common type of task used in competitions is a predictive task, in which one target variable must be predicted.

Proof-of-Learning

There are three types of actors involved in this consensus algorithm:

- *Suppliers*, which continuously propose new ML problems to the network, together with a data set.
- *Trainers*, which train their models on the proposed data set. They want to win the price linked to the competition. They can only use a fixed set of ML libraries.
- *Validators*, which are nodes of the blockchain chosen randomly. They propose the new blocks and evaluate the models on the test data.

Proof-of-Learning

Every time a new block is added to the blockchain, a ML problem must be solved.

- Suppliers publish the train dataset on the IPFS website and then inserts into the blockchain the hash of the dataset.
- Suppliers send the reward and a fee to an empty address.
- Every trainer sends a transaction containing the hash of the model to the blockchain, together with a fee.
- The supplier uploads the testing data set and then the trainer publishes its whole model on the IPFS.

Proof-of-Learning

- The chosen validators test the models on the testing data set.
- Validators reach consensus on the new block and receive the minted blockchain cryptocurrency plus the supplier's fee.
- The winning trainer receives the supplier's reward.
- The best model is saved off-chain.

How can the protocol ensure there will always be a number of tasks higher than the number of blocks which have to be created?

Conclusion

- ML and blockchains can efficiently cooperate to improve both technologies.
- It has been possible to detect illicit behaviours on the blockchain thanks to the Machine Learning.
- On the other way around, blockchains can remove the centralization problem typical of datasets.
- New results are expected in the next years as blockchain data and popularity grow.

References I

- Hao Hua Sun Yin et Al, Regulating Cryptocurrencies: A Supervised Machine Learning Approach to De-Anonymizing the Bitcoin Blockchain.
- Hao Hua Sun Yin et Al, A first estimation of cybercriminal entities in the Bitcoin ecosystem using supervised machine learning.
- Laura Alessandretti, Abeer ElBahrawy, LucaMaria Aiello, and Andrea Baronchelli, Anticipating Cryptocurrency Prices Using Machine Learning.
- Salim Lahmiri, Stelios Bekiros, Cryptocurrency forecasting with deep learning chaotic neural networks.
- Felipe Bravo-Marquez et Al, Proof-of-Learning: a Blockchain Consensus Mechanism based on Machine Learning Competitions.

References II

- Cutey G. Akcora et Al, PAKDD: Forecasting Bitcoin price with graph chainlets.
- Wei Xiong, Li Xiong, Smart Contract Based Data Trading Mode Using Blockchain and Machine Learning.
- Felipe Bravo-Marquez et Al, Proof-of-Learning: a Blockchain Consensus Mechanism based on Machine Learning Competitions.
- Wei Xiong, Li Xiong, Smart Contract Based Data Trading Mode Using Blockchain and Machine Learning.
- Fang Chen et Al, Machine Learning in/for Blockchain: Future and Challenges.
- Siddhi Velankar, Sakshi Valecha, Shreya Maji, Bitcoin Price Prediction using Machine Learning.