UNIVERSITÀ DI PARMA

Dipartimento di Ingegneria e Architettura

Corso di Laurea Magistrale in Ingegneria Informatica

Analisi Multivariata per la Previsione del Prezzo del Bitcoin con Tecniche di Machine Learning

Multivariate Data Analysis for Bitcoin Price Prediction Based on Machine Learning

Tesi di Laurea di:
Stefano Cavalli

Bitcoin è un sistema peer-to-peer ideato nel 2009 da una persona, o un gruppo di persone, sotto lo pseudonimo di Satoshi Nakamoto. È basato su tecnologia blockchain: una struttura dati organizzata a blocchi tra loro connessi attraverso un meccanismo crittografico, in grado di garantire, alle informazioni in essa contenute, veridicità, stabilità e immutabilità. Si è diffuso principalmente come metodo di pagamento per il Deep Web ma la sua natura altamente volatile ha attratto trader da tutto il mondo che si cimentano ogni giorno nello studio di metodologie per prevederne l'andamento.

Lo scopo di questo lavoro di Tesi è quello di creare un modello predittivo, grazie all'utilizzo dell'intelligenza artificiale, per cercare di stimare se in futuro il prezzo del bitcoin sarà maggiore o minore di quello attuale. Sono stati sviluppati diversi sistemi per l'acquisizione dei dati (di varia natura), necessari all'addestramento del modello di machine learning.

La prima feature tenuta in considerazione è quella relativa al sentimento espresso dagli utenti su Twitter. Si è scelto di realizzare un'infrastruttura che preveda l'utilizzo di un nodo *master* e molti nodi *slave*, in grado di poter scaricare un qualsiasi numero di tweet contentenenti una parola (o frase), in un range temporale definito. L'uso di un solo server non sarebbe stato possibile perchè dopo un tempo pari a circa 72 ore i processi smettono automaticamente di funzionare a causa del ban dell'indirizzo IP da parte di Twitter. È stato dunque necessario creare e configurare 22 server differenti che fungono da *slave* mentre un Virtual Private Server (VPS) ha la funzione di *master*. Il *master* tiene sempre traccia dei tweet già scaricati dal 1 gennaio 2009 al 15 febbraio 2019 e si mette in ascolto in attesa di uno *slave*, fino a che non ha ottenuto tutti i tweet preventivati inizialmente. Quando lo *slave* si connette, riceve un comando preciso dal *master* in cui vengono indicati l'intervallo temporale (giornaliero) entro la quale scaricare i tweet, il nome della parola che il tweet deve contenere (bitcoin) e la lingua (inglese). Dopo aver ricevuto il comando, lo *slave* invia al *master* i risultati ottenuti in formato CSV e cancella i file presenti in locale per liberare spazio. Dopodichè, lo *slave* ricomincia il processo, oppure termina la propria esecuzione.

Il *master* ha il compito di analizzare i tweet con un meccanismo di pre-processing prima di compiere un'operazione di sentiment analysis. Ogni tweet viene caricato nella tabella *dirty* presente sul database *tweets*, poi viene "pulito" attraverso la rimozione di quelle parole ritenute inutili per questo topic, perchè indici di azioni da parte di bot, e caricato in tabella *clean*. Si è scelto di utilizzare Vader come tool per la sentiment analysis. Ogni tweet viene analizzato e classificato secondo quattro differenti parametri: *positive_score*, *neutral_score*, *negative_score* e *compound*. I tweet vengono tra loro accorpati, considerandone il *timestamp*, e caricati in tabella *sentiment*. Ogni record corrisponde ad un giorno mentre i campi della tabella mostrano una media degli *score* e il numero totale di tweet per quel giorno.

La seconda fase di raccolta delle feature si concentra sull'analisi della blockchain di Bitcoin. Un full node di Bitcoin è stato installato e configurato sul VPS in modo da riuscire ad ottenere, in locale, tutti i blocchi della blockchain (operazione che ha richiesto circa sei giorni) che vengono

successivamente caricati nella tabella *blocks* del database *blockchain*. Electrs è un'implementazione in linguaggio Rust di un server Electrum e si occupa di mantenere la blockchain di Bitcoin indicizzata. Dopo averlo configurato e sincronizzato con il full node di Bitcoin è possibile farvi delle chiamate per ottenere informazioni sulle transazioni, utili a calcolare alcune feature. È stato creato un parser in grado di tradurre il codice binario presente nei blocchi in un formato testuale e leggibile. Per ciascuno degli oltre 600000 blocchi vengono effettuate delle chiamate in locale al server Electrs per capire quali sono gli input e gli output delle transazioni (la quantità di bitcoin spostati) e a quanto ammontano le commissioni. Tutte queste informazioni vengono accorpate per giorni e caricate nella tabella *days*.

L'ultima lista di feature comprende il valore storico del prezzo del bitcoin insieme a diversi indicatori finanziari. Attraverso il WebScraping effettuato sul sito web www.coinmarketcap.com, si ottengono tutti i dati storici relativi al bitcoin: apertura (O), chiusura (C), massimo (H), minimo (L) per ciascun giorno. Proprio a partire da queste informazioni, in maniera completamente dinamica viene creato un database contenente queste quattro informazioni e altri dieci indicatori finanziari.

Tutti i nuovi valori giornalieri delle feature descritte vengono ottenuti ogni giorno alle ore 01:00 in maniera automatica. A partire dai dati presenti nel database vengono creati dei file CSV a seconda dei parametri presi in considerazione. Si definisce:

$$\mathcal{B} = \{b_1, b_2, ..., b_{|B|}\} \tag{1}$$

come l'insieme dei valori di chiusura del prezzo del bitcoin, necessari a costruire le label del problema. n indica la quantità di valori da considerare per addestrare la rete neurale, si tratta dell'input. k rappresenta il numero di elementi da valutare per creare le label. z denota lo spostamento da compiere lungo B per poter calcolare i valori successivi del dataset, definito come segue:

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_{\left\lceil \frac{|B|-n-k+1}{z} \right\rceil}\} \tag{2}$$

Ogni elemento del dataset ha la seguente struttura:

$$\mathcal{D}_i = \{b_{1+(i-1)\cdot z}, ..., b_{n+(i-1)\cdot z}, l_i\}, \qquad i \in \{1, ..., |D|\} \tag{3}$$

dove l_i rappresenta la label utilizzata nel dataset:

$$\begin{cases} UP & se \quad \frac{\sum_{j=n+(i-1)\cdot z+1}^{n+(i-1)\cdot z+k} b_j}{k} \geq b_{n+(i-1)\cdot z} \\ DOWN & altrimenti \end{cases} \tag{4}$$

Si è scelto di utilizzare le Convolutional Neural Network (CNN) come modello di machine learning, in cui ogni operazione, che di norma viene effettuata in 2 dimensioni, qui avviene in maniera monodimensionale. In totale, le feature sono 17 e il dataset viene costruito a seconda dei valori di n, k e z che variano entro questi range: $n \in \{5, ..., 50\}$, $k \in \{2, ..., \lceil (\frac{n}{2}) \rceil + 1\}$, $z \in \{1, ..., n\}$. Il risultato mostra 18121 dataset (e addestramenti) differenti, con performance diverse tra loro. Anche il layer convolutivi sono diversi: la dimensione dei kernel varia in proporzione alla dimensione dell'input n. I risultati ottenuti in questa Tesi indicano che i parametri migliori da tenere in considerazione sono n=38, k=12, z=1 per ottenere un'accuratezza sul test del 74.24%.

In futuro, sarebbe interessante considerare un timeframe più ridotto, passando dalle ore ai giorni, ottenendo così molti più dati. Tuttavia, secondo diverse pubblicazioni scientifiche, aumentare la precisione di campionamento porta spesso un calo nelle prestazioni per quanto riguarda l'accuratezza sul data set. Un altro aspetto interessante potrebbe essere l'aggiunta arbitraria di nuovi indicatori finanziari e di altre feature come ad esempio il prezzo dell'oro o, più in generale, delle materie prime, dal momento che alcuni articoli parlano di legami tra il mercato delle commodity e quello delle criptovalute, mentre altri ancora ne indicano una correlazione con il mercato del Forex. Un'altra analisi che potrebbe portare a dei risultati soddisfacenti è quella di inserire, tra le feature già esistenti, il prezzo di altre criptovalute come Ethereum, Ripple, Bitcoin Cash e Bitcoin SV o valutare altre piattaforme di social media per la sentiment analysis.