

機械学習を用いた 暗号資産の収益率予測モデルの構築

于 海波 松永 諒

小澤 翔太

富山県立大学 情報システム工学科

2024 年 7 月 26 日

はじめに

近年、暗号資産市場は急速に拡大し、多くの投資家やトレーダーが参加している。この暗号資産市場は従来の金融市場とは異なる動きを見せ、価格の急激な変動や取引量の増減が顕著である。また、暗号資産取引所は 24 時間 365 日常に動いており、データ取得や取引を行うための公開 API を用意しているところがほとんどである。

よって、人間が市場を確認し取引の対応を行うよりもプログラムを用いて自動で取引を行うほうが、ほかの市場参加者に比べて有利になる部分がある。

本研究の目的

そこで本研究では、暗号資産取引所で公開されている過去の暗号資産の取引データや API を利用して、収益率予測機能を持つモデルの作成及び、そのモデルの評価を行う。

特徴量-RSI

RSI(Relative Strength Index, 相対力指数) とは, 金融市場で使用されるテクニカル分析の指標の一種であり, 特定の期間の値動きの強さを示す指標である. RSI は式 (1) で与えられる.

$$RSI = 100 - \frac{100}{1 + RS} \quad (1)$$

RS(Relative Strength) とは, 平均利得 (Average Gain) を平均損失 (Average Loss) で割ったものあり, 式 (2) で与えられる.

$$RS = \frac{AverageGain}{AverageLoss} \quad (2)$$

特徴量-EMA

EMA(Exponential Moving Average, 指数平滑移動平均) とは, 時系列データの平滑化やトレンドの抽出に使用される統計的な指標の 1 つである. EMA は式 (3) で与えられる.

$$EMA_t = \alpha X_t + (1 - \alpha) EMA_{t-1} \quad (3)$$

X_t は現在の時点 t での観測値で, EMA_t は時点 t での EMA の値である. α は平滑化のための係数であり, 通常 0 から 1 の間の値をとる.

EMA のメリットは, EMA は最新の価格データに, より高い重みを与えるため滑らかでトレンドをより正確に表示することにより, 変動する市場に適応しやすい.

しかし, これがデメリットでもあり急激な変動が起こった際に, それがノイズとして出現してしまいう可能性がある.

特徴量-SMA

SMA(Simple Moving Average, 単純移動平均) とは, 時系列データの平滑化やトレンドの把握に使われる統計的な指標の 1 つである. SMA は式 (4) で与えられる.

$$SMA_t = \frac{X_{t-n+1} + X_{t-n+2} + \cdots + X_{t-1} + X_t}{n} \quad (4)$$

SMA_t は時点 t での SMA の値であり, X_t は時点 t での観測値である. 期間が n である場合, データポイント数が n となる.

SMA のメリットは, 最新のデータと過去のデータを平等に扱うため, 安定している.

デメリットは, すべてのデータに同じ重みを与えるため, 急激な変動があった場合にその変動を捉えきれず, テンドの転換を遅れて捉えてしまう事がある.

概要

本システムはまず、暗号資産取引所から暗号資産のリアルタイムデータを取得してデータの前処理を行い、収益率予測機能を利用して収益率を予測する。その予測結果を基に暗号資産の売買を行い、それを繰り返すというシステムである。

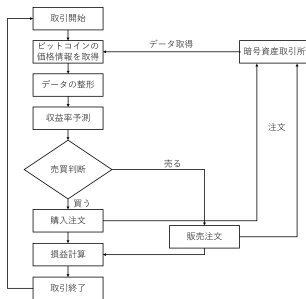


図 1: 暗号資産自動取引システムの全体像

過去データの取得

今回は GMO コインの API ドキュメントを参考にして、BTC の始値 (Open), 高値 (High), 低値 (Low), 終値 (Close), 出来高 (Volume) の 5 つの OHLCV の過去データを取得した。今回は 2023 年 10 月 2 日までの 100 日分の 1 分足のデータを取得した。

データの前処理

- ・ 目的変数の生成：10 分後の収益率を目的変数とする。
- ・ 特徴量の生成：BTC の始値, 高値, 低値, 終値, 出来高, RSI, SMA, EMA を特徴量として用いる。
- ・ 欠損値の穴埋め：直前の時間の値を利用して欠損値の穴埋めを行う。
- ・ データの正規化：データを 0 から 1 の範囲にする最小-最大スケールリングという方法を用いる。
- ・ データの分割：全データの 70 % をトレーニングデータ, 30 % をテストデータとする。

モデルの構造，学習

- ・ LSTM 層：3つの LSTM 層から構築されている．各 LSTM 層は 32 ユニットを持つ．
- ・ ドロップアウト層：各 LSTM 層の後に 2 層ある．ドロップアウト率は 0.2 に設定している．
- ・ 全結合層：ユニット数は 1 に設定している．
- ・ モデルのコンパイル：最適化アルゴリズムを Adam に設定し，学習率は 0.0005 である．
- ・ モデルの学習：今回，エポック数を 150，ミニバッチサイズを 64 に設定している．

評価の概要

評価指標として、R2, MAE, RMSE の3つを使用する.

R2

R2(R-squared) とは、統計学や機械学習などで使用される評価指標の1つであり、R2はモデルが目的変数の変動をどれだけ説明出来るかを示す. R2は式(5)で与えられる.

$$R2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

数式(5)において、各 y_i は実際の目的変数の値、 \hat{y}_i はモデルによる予測値、 \bar{y} は目的変数の平均値、 n はデータポイントの数を表す. 一般的に、 $0 \leq R2 \leq 1$ の範囲で評価され、数値が高いほどモデルの性能が高いことを示す.

R2 による評価結果

R2 による予測モデルの機能の評価を表 1 と表 2 で示す. 表 1 は GRU モデルの R2 の評価結果であり, 表 2 は LSTM モデルの R2 の評価結果である.

表 1: GRU モデルの R2

特徴量	訓練データ	テストデータ
なし	0.6749322925444785	0.6651717009689513
RSI	0.738622460018331	0.729981510177558
SMA	0.6955070654469848	0.6818013951883636
EMA	0.6804925596024708	0.6706496945695288
RSI SMA	0.782430750876208	0.776792664065447
RSI EMA	0.781083560370537	0.770992709180339
SMA EMA	0.683036049700632	0.6697838171951207
RSI SMA EMA	0.7508283973584714	0.7408636292074918

表 2: LSTM モデルの R2

特徴量	訓練データ	テストデータ
なし	0.7266783812254729	0.7172616257296095
RSI	0.770878586338471	0.7650348516411895
SMA	0.7239908181506745	0.7135706859935746
EMA	0.73885384742825	0.7291627314738351
RSI SMA	0.7959045696849647	0.782478303889052
RSI EMA	0.7906108779525182	0.7802905680805978
SMA EMA	0.7424865128989024	0.7313102216635474
RSI SMA EMA	0.7837917390137639	0.7702482020379458

MAE

MAE(Mean Absolute Error) とは、回帰モデルの性能を評価するための指標の 1 つである。MAE は予測値と実際の値との絶対値の平均を計算し、モデルの予測が実際のデータとどれだけ離れているのかを示す。MAE は式 (6) で与えられる。

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

数式 (6) において、 n はデータポイントの数を表し、 y_i は実際の目的変数の値、 \hat{y}_i はモデルによる予測値を表す。MAE の値は小さいほど良い値である。(ただし、MAE は 0 以上)

MAE による評価結果

MAE による予測モデルの機能の評価を表 3 と表 4 で示す。表 3 は GRU モデルの MAE の評価結果であり、表 4 は LSTM モデルの MAE の評価結果である。

表 3: GRU モデルの MAE

特徴量	訓練データ	テストデータ
なし	0.050227986160070355	0.05121246803902003
RSI	0.044045224578800564	0.04442748482031604
SMA	0.049416153475860355	0.04979394110662729
EMA	0.052575282920683045	0.053627303170900836
RSI SMA	0.04101977038441864	0.041180107072966876
RSI EMA	0.04149273756397863	0.04188118148380276
SMA EMA	0.05248781691391784	0.053348590394212456
RSI SMA EMA	0.04504523692404792	0.046043624752074395

表 4: LSTM モデルの MAE

特徴量	訓練データ	テストデータ
なし	0.04581846200215947	0.046739969770889955
RSI	0.0448039020541127	0.0455486349658181
SMA	0.04435256944600605	0.04524741272751368
EMA	0.04477239856000979	0.04562683819960324
RSI SMA	0.04063198222407326	0.04046958420563298
RSI EMA	0.04236332916196131	0.04322542356244045
SMA EMA	0.044597866617397654	0.04553965578779903
RSI SMA EMA	0.044539640297429602	0.04483013899748297

RMSE

RMSE(Root Mean Squared Error) とは、回帰モデルの性能を評価する指標の 1 つである。RMSE は予測値と実際の値との差の平方の平均を計算し、その後平方根をとることで、誤差の絶対値の平均を示す。RMSE は式 (7) で与えられる。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

数式 (7) において、 n はデータポイントの数を表し、 y_i は実際の目的変数の値、 \hat{y}_i はモデルによる予測値を表す。

RMSE による評価結果

RMSE による予測モデルの機能の評価を表 5 と表 6 で示す。表 5 は GRU モデルの RMSE の評価結果であり、表 6 は LSTM モデルの RMSE の評価結果である。

表 5: GRU モデルの RMSE

特微量	訓練データ	テストデータ
なし	0.08675475276067778	0.08102379390046484
RSI	0.07122767571808837	0.06787396519727097
SMA	0.08396435808456196	0.07798685678958711
EMA	0.08995691951096763	0.083938365058498
RSI SMA	0.06824213284287953	0.06730549572658496
RSI EMA	0.06878485107076622	0.06863413099436132
SMA EMA	0.08566655404182509	0.08046382487380561
RSI SMA EMA	0.07338204757382938	0.07147107946836288

表 6: LSTM モデルの RMSE

特微量	訓練データ	テストデータ
なし	0.07955058072044241	0.0744550011908708
RSI	0.07779299560097867	0.07276093121531924
SMA	0.07698996693190652	0.07227573218308717
EMA	0.07775855441434507	0.07287116441979366
RSI SMA	0.0684914703772196	0.06615394207853025
RSI EMA	0.06854928509979566	0.06700792216197486
SMA EMA	0.07721583143996898	0.07258168883703013
RSI SMA EMA	0.07531037964103964	0.07173913604204792

考察

今回は LSTM モデルの RSI, SMA の特徴を追加したモデルが最も評価が高かった. しかし, どのモデルも R^2 を見てみると 0.6 以上を示しており, 一定以上の正確性を持っていると言える.

また, 3つの評価指標を確認すると学習データとテストデータの大幅な乖離が起きていないため過学習をしていないことも確認できる. 以上より, 今回のモデルの構築は成功したと考えられる.

まとめ

- 本研究の目的は暗号資産の自動取引システムを作成し、市場参加者が常に監視し市場に張り付く必要がなく、暗号資産を売買できるようにすることである。
- 本研究ではそのシステムの中で、暗号資産の収益予測機能を構築し、3つの特徴量を組み合わせてモデルの評価やモデルの見直しを行った。
- 今回作成したモデルを今後更に改善することが重要だと考えられる。