ARIMA-GA-SVRによる株価予測モデル

A Hybrid ARIMA-GA-SVR Model for Stock Price Forecasting

卓　越[1*]　　森本　孝之[1,2]
Yue Zhuo[1]　Takayuki Morimoto[1,2]

[1] 関西学院大学理工研究科
[1] Graduate School of Science and Technology,Kwansei Gakuin University
[2] 関西学院大学理学部
[2] School of Science,Kwansei Gakuin University

Abstract: Autoregressive integrated moving average (ARIMA) is a widely used linear model with great performance for time series forecasting problems. Supplemented by support vector regression (SVR), an effective method to solve the nonlinear problem with a kernel function, ARIMA-SVR model captures both linear and nonlinear patterns in stock price forecasting. However, it does not have high accuracy and parameter selection speed when its parameters are chosen by the traditional method. Therefore, in this study, we applied genetic algorithm (GA) to optimize the parameter selection process of SVR to improve the performance of the ARIMA-SVR model. Subsequently, we built the ARIMA-GA-SVR model by integrating ARIMA with optimized SVR. Finally, we used actual stock price data to compare the forecasting accuracy of the proposed model, ARIMA and ARIMA-SVR models using error functions. The result shows that the proposed ARIMA-GA-SVR model outperforms other models.

1　はじめに

長期的な観点から, ある企業の株価は企業の業績や財務状況等によって動いている.Huang[8] は業績や財務状況などの指標を用い,GA により指標の選択を行うことと SVR のパラメーターを最適するハイブリッドモデルで中長期の株価を予測し,SVR はファンダメンタルズに基づく株価予測に対する有効であり,GA は SVR モデルの性能を向上させたことを示した. しかし, 短期的な動きを予測することはそれらのファクターで説明することは困難である.

そのため, 株価の短期的な予測において, 日次の株価を用いる時系列分析方法が広く用いられている. この中に ARIMA モデルは主要な方法の一つである.ARIMA モデルとは, 今期の株価を過去の株価で近似する AutoRegressive（AR）モデルと過去のランダム項を移動加重する Moving Average（MA）モデルを組み合わせ, データの非定常性を差分で解消するモデルである.2005 年に Pai ら [11] は株価を線形部分と非線形部分に分け, それぞれ ARIMA モデルと SVR モデルで予測する手法を提案した.Adebiyi ら [2] は株価の日次動きの定常性を判断した上,ARIMA モデルで株価の予測を行った. 彼らの研究により,ARIMA モデルの株価予測問題に対する有効であることと SVR モデルは ARIMA モデルを改善できることが示されている. しかし,SVR モデルはパラメーターに対し, 非常に敏感であるため, パラメーターの小さな変化でもモデルのパーフォマンスに対する大きな影響を与える.SVR モデルのパラメーターを決めることは困難である.

本研究の目的は,GA を導入し,SVR モデルのパラメーターをより精確的に選択することを行い, 改良されたモデルを米国株式市場において, 既存の ARIMA モデル及び ARIMA-SVR モデルとの比較を行う. 実証分析の結果として, 提案モデルはより良いパーフォマンス及びより速い計算速度を持つことを示す.

*連絡先： 関西学院大学理工研究科
〒 669-1330 兵庫県三田市学園上ケ原 1 番
E-mail:yue-zhuo@kwansei.ac.jp

2 方法

2.1 ARIMA(Autoregressive Integrated Moving Average)

ARIMA(p,d,q) モデルは最初 1970 に Box ら [3] に提案された線形非定常性を d 階差分で解消した自己回帰移動平均モデルである. このモデルは非常に古典的な時系列モデルであり, 特に金融予測領域においてよく用いられている.

ARIMA(p,d,q) モデルには, 時刻 t の観測値が過去の観測値と過去の誤差項の線型結合

$$Y_t = \phi_0+\phi_1 Y_{t-1}+...+\phi_p Y_{t-p}+\epsilon_t-\theta_1\epsilon_{t-1}-...-\theta_q\epsilon_{t-q} \tag{1}$$

により表される. ここに Y_tは観測値, ϵ_tは誤差項, $\hat{Y}_t$を予測値に仮定すると, 誤差項ϵ_tは $Y_t-\hat{Y}_t$になる.ϕ_iとθ_jは係数, p と q は自己回帰と移動平均の階数である.

例え,ARIMA$(1,0,1)$ は

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t - \theta_1\epsilon_{t-1}$$

である. しかし,ARIMA モデルは線形モデルのため, 非線形データに対する処理能力は制限されている.

2.2 SVR(Support Vector Regression)

SVR モデルは線形の回帰分析手法だが, 説明変数を高次元への写像を通じて, 非線形回帰モデルに変換することができる. モデルは

$$\hat{y}_i = \boldsymbol{\omega}\phi(\boldsymbol{x_i}) + b \tag{2}$$

ここに $\phi(\cdot)$ は高次元へ変換する写像である. 以下の方法より, 二つの係数 $\boldsymbol{\omega}$ と b を確定できる.

$$\min R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\epsilon(d_i,y_i) + \frac{1}{2}\|\omega\|^2 \tag{3}$$

$$\min L_\epsilon(\hat{y}_i,y_i) = \begin{cases} |\hat{y}_i-y_i|-\epsilon, & |\hat{y}_i-y_i|\geq\epsilon, \\ 0, & others, \end{cases} \tag{4}$$

ここに $\hat{y}_i$は i 時刻の予測値, y_iは i 時刻の観測値,L_ϵ は感度係数 ϵ の損失関数,C は誤差を許容する程度を表す罰則係数である. そして, スラック変数 ξ,ξ^* を導入し, 式 (3) を以下の式に変換する.

$$\min R(\boldsymbol{\omega},\xi,\xi^*) = \frac{1}{2}\|\omega\|^2 + C^*(\sum_{i=1}^{N}(\xi_i+\xi_i^*)) \tag{5}$$

束縛条件は

$$\hat{y}_i - y_i \leq \epsilon + \xi_i^*,$$

$$y_i - \hat{y}_i \leq \epsilon + \xi_i,$$

$$\xi_i,\xi_i^* \geq 0, i = 1,2,3,...,N$$

である. ここにラグランジュ乗数 α と $\hat{\alpha}$ を導入し, 双対問題に変換する.

$$\begin{aligned} D(\boldsymbol{\omega},b,\alpha,\hat{\alpha},\xi,\xi^*,\mu,\hat{\mu}) = {} & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i+{\xi^*}_i) \\ & - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\hat{\mu}_i{\xi^*}_i \\ & + \sum_{i=1}^{N}\alpha(\hat{y}_i - y_i - \epsilon - \xi_i^*) \\ & + \sum_{i=1}^{N}\hat{\alpha}(y_i - \hat{y}_i - \epsilon - \xi_i) \end{aligned} \tag{6}$$

最後に, 回帰方程式の解は,

$$f(\boldsymbol{x},\alpha,\hat{\alpha}) = \sum_{i=1}^{N}(\alpha_i-\hat{\alpha}_i)\phi(\boldsymbol{x})\phi(\boldsymbol{x_i}) + b \tag{7}$$

である. しかし, ここには高次元計算の $\phi(\boldsymbol{x})\phi(\boldsymbol{x_i})$ があり, 計算を簡単にするため, カーネル関数 $k(\boldsymbol{x},\boldsymbol{x_i}) = \phi(\boldsymbol{x})\phi(\boldsymbol{x_i})$ を導入する. 本研究では, ガウシアンカーネル関数

$$k(\boldsymbol{x},\boldsymbol{x_i}) = exp(-\|\boldsymbol{x}-\boldsymbol{x_i}\|^2/2\sigma^2)$$

を使用する. この中 $\sigma^2=\gamma$ である.

2.3 GA(Genetic Algorithms)

遺伝的アルゴリズム（GA）は 1975 年に Holland が提案し, 生物進化の過程をシミュレーションする最適化アルゴリズムの一つである. 本研究には,GA を用いて SVR モデルのパラメーターを最適化する.

GA の手順は、以下 (図 1) のようである. 初期集団の生成は, 決定された個数と長さの染色体（解の候補）をランダムに生成する. 初期集団が生成された後, 各個体の適応度を適応関数で評価する. 評価すると, 高い適

用語	意味
染色体	解の候補
染色体長さ	解の個数
各世代染色体の個数	解の候補の個数
評価関数	解の適応度を評価する関数
最大世代数	アルゴリズムの反復次数
選択	適応度低い解を淘汰する
変異と交差	新しい候補を作る

表 1: GA に必要なエレメントリスト

応度の個体を選び, 終了条件が満たされていない際に, 交差と変異の操作を行い, 次の世代を生成する. そして, 終了条件が満たされるまで, 以上の操作を繰り返す.

本研究では, 最適化するパラメーターは三つがあるため, 染色体の長さを 3 に設定する. そして, 適応関数は SVR モデルのインサンプル予測結果の MAE を使用する. 以上の流れを 100 世代まで実行する.

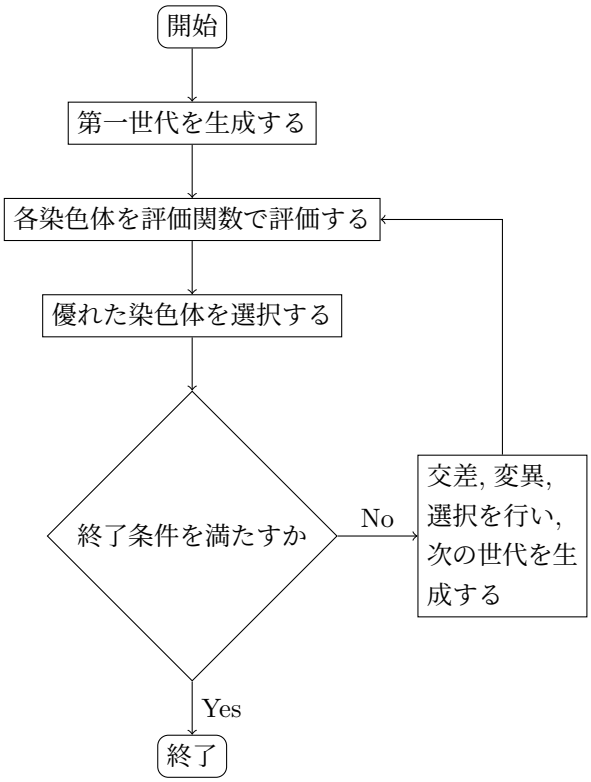

図 1: GA 流れ図

2.4 提案手法

株価 S_t は線形パターンと非線形パターンに分けることが可能と仮定すると, 株価は

$$S_t = Y_t + N_t \tag{8}$$

により表される. ここに,Y_t は線形パターン,N_t は非線形パターンである. そして,ARIMA モデルの線形パターンの予測値は $\hat{Y}_t$ により表すと, 残差 ϵ は $\epsilon = S_t - \hat{Y}_t$ である.

そこに,SVR モデルの非線形パターンの予測値 $\hat{N}_t$ を表すと

$$\epsilon = \hat{N}_t + \delta_t \tag{9}$$

δ_t はランダム誤差である. 従って, 株価の予測値 $\hat{S}_t$ は

$$\hat{S}_t = \hat{Y}_t + \hat{N}_t \tag{10}$$

である.

具体的に, 提案手法の手順は

1. ADF Test と KPSS Test でデータの定常性を判断する.
2. ARIMA(p, d, q) で, 株価を予測する.
3. ステップ 2 の ARIMA モデルにより, 残差データを計算する.
4. 残差データを SVR モデルに入れ,GA で SVR モデルの最適化を行う.
5. 株価の予測値を計算する.
6. ARIMA-GA-SVR モデルの検証期間の誤差関数 MAE,MSE,MAPE を計算する.
7. ARIMA のパラメーターを変更すると, ステップ 2 から繰り返す.

である. 終了条件を満たすと, 全てのモデルの中に, 検証期間で誤差関数 MAE が最も低いモデルを選択する.

3 実証分析

実証分析において, 提案モデルの性能を評価するため, 提案モデルと ARIMA モデルと ARIMA-SVR モデルの比較を行う.SVR モデルと ARIMA モデルの実行は Python の Scikit-learn[12] と Statsmodels[13] で実現する. そして, 実証分析のデータセットの分け方は,90%は訓練期間,5%は検証期間,5%はテスト期間である. データは米国株式市場の四つの銘柄 (表 2) の 2021/06/04 から 2022/04/01 まで（表 3）の日次終値である.

銘柄	コード
Apple	AAPL
NVIDIA	NVDA
Microsoft	MSFT
Tesla	TSLA

表 2: 銘柄&コード

Training Set	2021/06/04-2022/03/03
Validation Set	2022/03/04-2022/03/01
Test Set	2022/03/18-2022/04/01

表 3: データセット

データの定常性検定は表 4と表 5によってあらゆる銘柄のデータは定常ではないことがわかっており, 差分すると, データは定常になることがわかった. 従って,ARIMA モデルの d は 1 にする.

銘柄	元データ		一階差分	
	ADF Test	KPSS Test	ADF Test	KPSS Test
Apple	-1.933	1.709	-12.748	0.0879
NVIDIA	-1.916	1.152	-11.298	0.103
Microsoft	-2.356	0.900	-14.958	0.241
Tesla	-1.498	1.323	-13.774	0.076

表 4: 検定統計量

有意水準	ADF value	KPSS value
10%	-2.574	0.347
5%	-2.876	0.463
1%	-3.462	0.739

表 5: 臨界値

そして,ARIMA モデルと ARIMA-SVR モデルはグリッドサーチで最適なパラメーターを検索する. この中,ARIMA モデルの基準は赤池情報量基準,ARIMA-SVR モデルの基準は平均標準誤差である. 各モデルは ADFtest と KPSS で d を判断した上, 他のパラメーターは表 6のようになる. それ以外, 遺伝的アルゴリズムの最大世代数は 100 世代に設定する. 実証分析の結果は表 7と図 2435で表示されている. 結果から見ると, 提案した ARIMA-GA-SVR モデルは三つの誤差関数指標

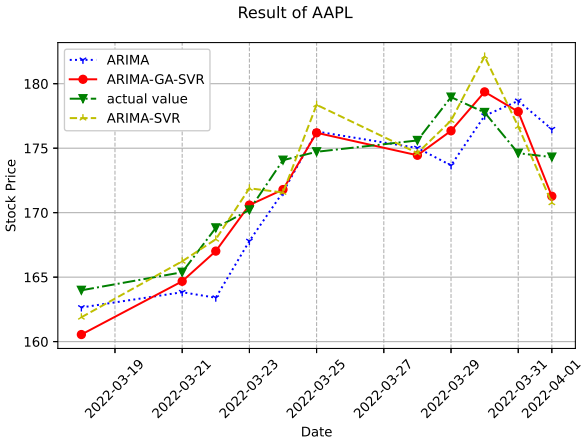

図 2: Apple 予測結果

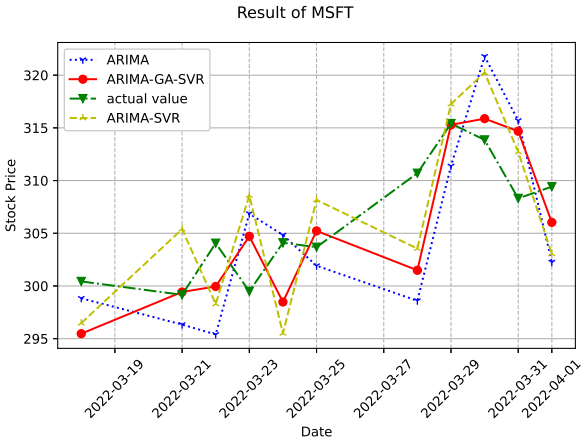

図 3: Microsoft 予測結果

の元に, 他のモデルより優れたパーフォマンスを示した.ARIMA-SVR モデルは提案モデルと同じ精度を持っているはずだが, 従来の方法でモデルのパラメーターを最適化し,SVR モデルの精度を向上させることは困難である. その原因は, パラメーターの範囲は離散的であり, 精度が高くない一方,GA は連続区間において探索することができる.

4 まとめ

本研究は,ARIMA モデルとハイブリッドモデル ARIMA-GA-SVR モデルの One-step 株価予測性能の比較を行った.GA 方法は連続区間において探索プロセスを行えるため, 最適なパラメーターを選ぶことができた. そして, 提案モデルは線形部分と非線形部分両方を捕まえられ, モデルの精度を向上させ, 四つの銘柄のデータセットの中, 提案モデルは最も優れたパーフォマンスを示した.

今後の課題として

	ARIMA	ARIMA-SVR	ARIMA-GA-SVR
$p \in Z$	$0 \leq p \leq 7$	$0 \leq p \leq 7$	$0 \leq p \leq 7$
$q \in Z$	$0 \leq q \leq 7$	$0 \leq q \leq 7$	$0 \leq q \leq 7$
γ	-	$\{2 \times 10^{-5}, 2 \times 10^{-4}, 2 \times 10^{-3}, 2 \times 10^{-2}, 2 \times 10^{-1}, 2, 20\}$	$[10^{-6}, 10^{2}]$
C	-	$\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{-2}, 10^{-3}\}$	$[10^{-2}, 2 \times 10^{2}]$
ϵ	-	$\{10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$	$[10^{-6}, 10^{2}]$

表 6: パラメーター範囲

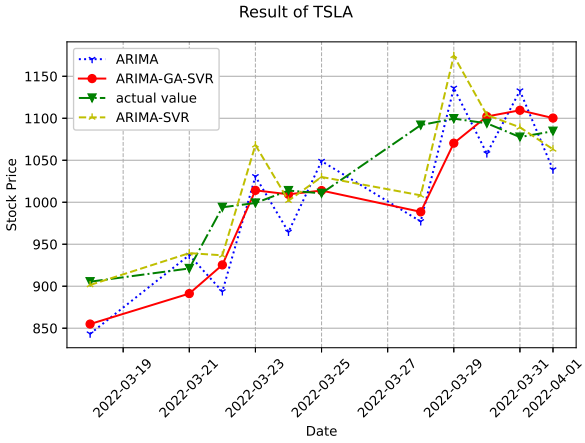

図 4: Tesla 予測結果

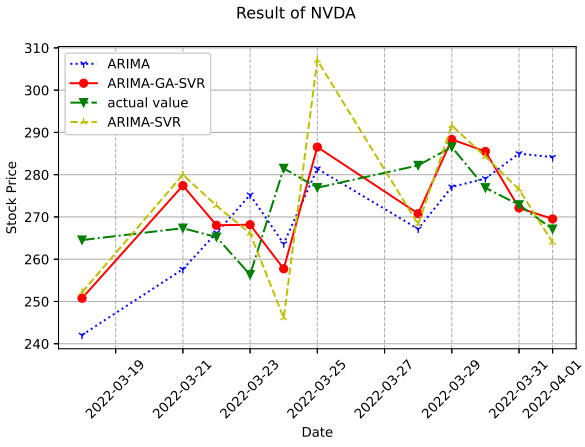

図 5: NVIDIA 予測結果

(a) ARIMA Model

	MSE	MAE	MAPE
AAPL	8·861233	2·464079	0·014427
NVDA	186·519716	11·836392	0·044284
MSFT	43·142881	5·593744	0·018325
TSLA	3636·218629	53·250266	0·054322

(b) ARIMA-SVR Model

	MSE	MAE	MAPE
AAPL	6·208698	2·220098	0·012760
NVDA	265·678176	12·814605	0·047332
MSFT	38·171782	5·845108	0·019150
TSLA	2017·319194	34·662781	0·033345

(c) ARIMA-GA-SVR Model

	MSE	MAE	MAPE
AAPL	4·845554	1·971050	0·011492
NVDA	119·083661	8·797943	0·032871
MSFT	22·163092	3·893828	0·012822
TSLA	1931·663115	32·723025	0·033758

表 7: 各モデルの誤差関数の比較

1. 本研究では最適な自己回帰と誤差項の階数の探索を行わず, 直接 0 から 7 までの整数に制限した. 将来は時系列モデルの分析を行い, モデルに対して最適な階数の探索を行う.

2. 今回は, 単変量モデルの予測パーフォマンスを検証した. 将来は多変量に対して共分散を考え, 多変量に対する効果を検証する.

参考文献

[1] Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. :comparison of arima and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*, Vol. 2014, , 2014.

[2] Adebiyi A Ariyo, Adewumi O Adewumi, and Charles K Ayo. :stock price prediction using the arima model. In *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*, pp. 106–112. IEEE, 2014.

[3] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *:Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[4] Kuan-Yu Chen and Cheng-Hua Wang. :support vector regression with genetic algorithms in forecasting tourism demand. *Tourism management*, Vol. 28, No. 1, pp. 215–226, 2007.

[5] Ching-Hsue Cheng and Huei-Yuan Shiu. A novel ga-svr time series model based on selected indicators method for forecasting stock price. In *2014 International Conference on Information Science, Electronics and Electrical Engineering*, Vol. 1, pp. 395–399, 2014.

[6] Joao Fausto Lorenzato de Oliveira and Teresa B Ludermir. :a distributed pso-arima-svr hybrid system for time series forecasting. In *2014 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 3867–3872. IEEE, 2014.

[7] Ahmed Fawzy Gad. :pygad: An intuitive genetic algorithm python library. *arXiv preprint arXiv:2106.06158*, 2021.

[8] Chien-Feng Huang. :a hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, Vol. 12, No. 2, pp. 807–818, 2012.

[9] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. :testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, Vol. 54, No. 1-3, pp. 159–178, 1992.

[10] James G MacKinnon. :critical values for cointegration tests. Technical report, Queen's Economics Department Working Paper, 2010.

[11] Ping-Feng Pai and Chih-Sheng Lin. :a hybrid arima and support vector machines model in stock price forecasting. *Omega*, Vol. 33, No. 6, pp. 497–505, 2005.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. :scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.

[13] Skipper Seabold and Josef Perktold. :statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[14] UVBR Thissen, R Van Brakel, AP De Weijer, WJ Melssen, and LMC Buydens. :using support vector machines for time series prediction. *Chemometrics and intelligent laboratory systems*, Vol. 69, No. 1-2, pp. 35–49, 2003.

[15] 高野祐一. : サポートベクトルマシンとカーネル法 (特集 機械学習の手法と役割期待). オペレーションズ・リサーチ= Communications of the Operations Research Society of Japan: 経営の科学, Vol. 65, No. 6, pp. 304–309, 2020.

[16] 森健一. :arima モデルによる需要予測. 日本経営工学会誌, Vol. 26, No. 4, pp. 313–319, 1976.

[17] 中川慧, 今村光良, 吉田健一. : 株価変動パターンの類似性を用いた株価予測. 人工知能学会全国大会論文集 第 31 回 (2017), pp. 2D11–2D11. 一般社団法人 人工知能学会, 2017.

[18] 北野宏明. : 遺伝的アルゴリズム. 人工知能, Vol. 7, No. 1, pp. 26–37, 1992.