ISSN: 2688-9323

The Research on Optimal Investment Strategy based on Apriori ARIMA Model

Jiwen Liu*

School of Instrument and Electronics, Zhongbei University, Taiyuan, Shanxi, 030000, China

Abstract

Market traders buy and sell volatile assets frequently, with a goal to maximize their total return. There is usually a commission for each purchase and sale. This paper focuses on how to design a quantitative trading strategy to optimize the investment strategy, and then design the ARIMA model based on the Apriori algorithm with four exponential functions to ope with different days of rise or fall respectively. And the model is flexible in its response to the market and can make better feedback in the current market environment. Besides, the model can accurately analyze the relationship between the number of days of increase and the return, and the model in the process of running to add or reduce positions, the principal does not need to be taken out or deposited at once, which can effectively reduce the riskiness of the model.

Keywords

Aprior; Fitting Function; ARIMA; Multi-cycle Integrated Forecasting.

1. Introduction

Market traders buy and sell volatile assets frequently, with a goal to maximize their total return. There is usually a commission for each purchase and sale. In this problem we analyze two financial products, gold and bitcoin, with the goal of maximizing the returns that may be brought to us by this volatile asset.

This paper used the data of gold and Bitcoin to design a quantitative trading model to optimize the investment strategy, and perform a sensitivity analysis to evaluate the pros and cons of the model.

2. Model Building and Results

2.1. The Price Trend Forecasting Model

2.1.1. Model Building

Smoothing treatment: ADF is a commonly used unit root test, his original hypothesis is that the series has a unit root, i.e., non-smooth, for a smooth time series data, it is necessary to be significant at a given confidence level and reject the original hypothesis.

Using differencing to convert serial data into balanced series, differencing can convert data into a smooth series. First-order differencing refers to the subtraction operation between two series values whose original series values are one period apart; K-order differencing is the subtraction between two series values that are K periods apart. If a time series has smoothness after the difference operation, the series is a differential smooth series and can be analyzed using ARIMA model.

After determining the unstable, the differential is performed in order of 1st order, 2nd order, 3rd order...until it is smooth.

Randomization treatment, for a purely random sequence, there is no correlation between the values of the sequence, the sequence is undergoing completely unordered random fluctuations,

ISSN: 2688-9323

and the analysis of the sequence can be terminated. Usually, a linear model is built to fit the development of the sequence, by which useful information about the sequence is extracted. the ARMA model is the most commonly used model for fitting smooth sequences.

A time series is pre-processed and determined to be a smooth non-white noise series, and then time series modeling can be performed.

Here, we use ARMA (p, q) series, if $\{X_t, t = 0, \pm1, \pm2 \cdots\}$ is a zero-mean smooth series that satisfies the following model,

$$X_t - \phi_1 X_{t-1} - L - \phi_p X_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - L - \theta_q \varepsilon_{t-q} \quad (1)$$

Among them, ε_t is a smooth series with zero mean and variance is σ_ε^2. Then X_t is said to be an autoregres sliding average series of order p, q, abbreviated as a ARMA (p, q) series.

Applying the operator polynomial $\phi(B), \theta(B)$, equation (1) can be written as,

$$\phi(B)X_t = \theta(B)\varepsilon_t \quad (2)$$

For a general smooth series $\{X_t, t = 0, \pm1, \pm2, L\}$, let its mean value $E(X_t) = \mu$, satisfy the following model.

$$(X_t - \mu) - \phi_1(X_{t-1} - \mu) - L - \phi_p(X_{t-p} - \mu) = \varepsilon_t - \theta_1 \varepsilon_{t-1} - L - \theta_q \varepsilon_{t-q} \quad (3)$$

Among them, ε_t is a smooth series with zero mean and variance is σ_ε^2. Using the backward shift operator $\phi(B), \theta(B)$, equation (3) can be tabulated as,

$$\phi(B)(X_t - \mu) = \theta(B)\varepsilon_t \quad (4)$$

With respect to the operator polynomial $\phi(B), \theta(B)$, the following additional assumptions are usually made.

1)$\phi(B)$ and $\theta(B)$ don't have public factor, at the same time $\phi_p \neq 0, \theta_q \neq 0$.

2) The roots of $\phi(B) = 0$ are all outside the unit circle, a condition known as the smoothness condition of the model.

3) The condition that all the roots of $\theta(B) = 0$ are outside the unit circle is called the reversibility condition of the model

2.1.2. Results

From the above analysis, for getting the daily investment amount of gold and bitcoin, we use a time series model to analyze the given historical data and predict the price of the financial product for the next trading day by using the historical data to give the best strategy.

Let the forecasted gold price for the next trading day be $\overline{X_{t+1}}$,and the actual price on that day is X_{t+1} ,then the relative error δ_t can be expressed as:

$$\delta_t = \frac{|\overline{X_{t+1}} - X_{t+1}|}{X_{t+1}} \times 100\% \quad (5)$$

To test whether the time series model is smooth, view the results of the ADF test and analyze whether it can significantly reject the hypothesis that the series is not smooth based on the analyzed t-value ($p < 0.05$ or 0.01);

ISSN: 2688-9323

Table 1. ADF Inspection Form

Variables	Difference order	t	P	AIC	Threshold		
					1%	5%	10%
	0	-0.434	0.904	9957.828	-3.436	-2.864	-2.568
USD (PM)	1	-8.159	0.000***	9948.534	-3.436	-2.864	-2.568
	2	-12.877	0.000***	9993.297	-3.436	-2.864	-2.568

Note: *** represent 1% significance levels, respectively.

The original data plot, model fitted values, and model predicted values of this time series model are shown below.

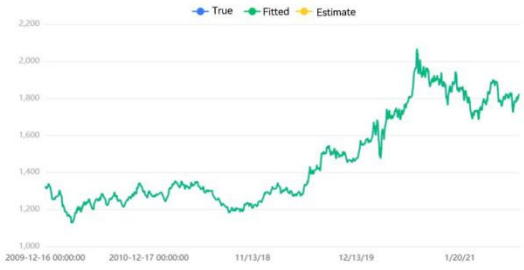

Figure 1. Time Series Chart

Figure 2. Time Series Chart (Partial)

With the two figures above, it is easy to see that the time series model we used has a good response to price forecasting, which further validates the accuracy of our forecasts using the model.

R^2 represents the degree of fit of the time series, and the closer to 1 the better the effect. The goodness-of-fit $R^2 = 0.997$ is obtained by statistical calculation.

Gold and Bitcoin forecast results for the next 5 days are shown in the table below.

Table 2. Price Forecast

Assets	1	2	3	4	5
Gold	1794.975	1795.35	1795.724	1796.099	1796.474
Bitcoin	46263.824	46327.72	46352.807	46377.894	46402.891

By forecasting two financial products, gold and bitcoin, using a time series model, we get the predicted prices of the products for five trading days as above, for which we can determine the investment direction by the predicted prices.

2.1.3. Optimized Price Forecasting Model

In order to make the forecast results more accurate, according to the principle of the forecasting model mentioned above, it is known that the forecast of future financial product prices is based on the changes in historical prices, and the pattern of price changes has a lot to do with the way of data statistics, we use multi-period integrated forecasting.

Table 3. Price Forecast (Optimization)

Periodicity	$t_1 = 1$	$t_2 = 2$	$t_3 = 3$	$t_4 = 4$	$t_5 = 5$
Gold	1802.91	1804.58	1805.42	1845.36	1825.87
Bitcoin	45627.01	45213.21	49007.78	53574.77	50897.99

ISSN: 2688-9323

Therefore, the historical prices of gold and bitcoin are processed and forecasted, taking the price forecast of one of the cycles as an example.

The relative error between the optimized model and the actual value is calculated and compared with the relative error of the original model, and it is found that the relative error of the optimized model is smaller, so the model established above is valid for it.

2.2. Trading Strategy Model

2.2.1. Model Building

The historical data is analyzed to get the rise and fall of gold and bitcoin. At this point, with a large amount of historical data, we analyze the pattern of rising and falling trading prices, and it is reasonable to assume that we can sell when the rise exceeds 50% of the historical rise, and similarly it is reasonable to assume that we can buy the product when the fall exceeds 50% of the historical fall. We then analyze the data using the Aprior algorithm.

Affiliation Rules: An association rule is an implication expression shaped as $X \rightarrow Y$, where X and Y are disjoint sets of terms, i.e., $X \cap Y = \emptyset$. The strength of an association rule can be measured by its support and confidence.

Support:

$$\text{Support}(X, Y) = P(X, Y) = \frac{\text{num}(xy)}{\text{num}(\text{all samples})} \tag{6}$$

Confidence:

$$\text{Cinfidence}(XY) = P(x|Y) = \frac{P(xy)}{P(y)} \tag{7}$$

Minsupport, the minsupport is the artificially specified threshold that indicates the minimum importance of the item set in statistical significance.

Minconfidence, the minconfidence is also an artificially specified threshold that indicates the minimum reliability of an association rule. This association rule is called strong only if the support and confidence reach both minimum support and minimum confidence.

Frequent itemset, the set of all items that satisfy the minimum support.

Connection step: the self-connection of Frequent (k-1) item sets L_{K-1} generates candidate k item sets C_K.

And then use the Pruning strategy and deletion strategy.

Once the frequent itemsets are identified, strong association rules can be generated directly from them.

The generation steps are as follows,

(1) For each frequent itemset itemset, generate all non-empty subsets of itemset (these non-empty subsets must be frequent itemsets).

(2) For each non-empty subset s of itemset, if

$$\frac{\text{sup port\_count(t)}}{\text{sup port\_count(t)}} \geq \text{min\_conf} \tag{8}$$

Then the output *s* to *l-s*, where min_conf is the minimum confidence threshold.

Algorithm flow,

(1) First perform a scan of the database and count the number of occurrences of each item to form a candidate 1-item set.

(2) Filtering frequent 1-item sets based on the minsupport threshold.

(3) Combine frequent 1-item sets to form candidate 2-item sets.

(4) A second scan of the database, counting for each candidate 2-item set and filtering for

ISSN: 2688-9323

frequent 2-item sets.
(5) Repeat the above process until the candidate set is empty.
(6) Based on the generated set of frequent items, the management rules are generated by calculating the corresponding confidence levels.
Where if the itemset contains K different items, it is called K-itemset. The candidate K-item set is denoted as C_K, the frequent K-item set is denoted as L_K.

2.2.2. Results

Using gold as the result example, we use MATLAB to statistically analyze the data and calculate the actual rise and fall of the financial product for each trading day.

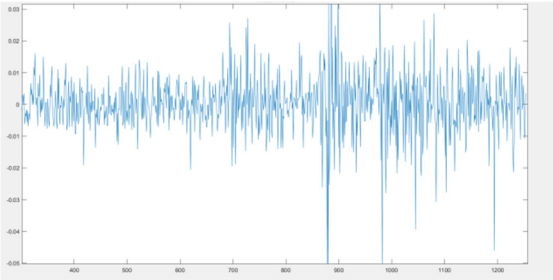

Figure 3. Gold up and down

After calculating their ups and downs on the trading day, find their median ups and downs respectively. The Median of the increase $M_{0.5}$ is 0.45%, and the $M_{0.5}$ is -0.14%.
At this point, the Apriori is introduced, and the subset of 2, 3, 4 and 5 consecutive rises or falls is found through the function of discovering frequent item sets of the Apriori. By analyzing the pattern of multiple consecutive rises and falls, the best point of time to invest and throw out is found, and the number of subsets of historical transactions in which the phenomenon will occur when 2, 3, 4, 5, consecutive rises or falls are analyzed here.
The following table is obtained by setting the continuous rise as m_x,and the continuous fall as n_x,

Table 4. The number of subsets of gold continuous up and down

Number of consecutive rises	m_x	Number of consecutive declines	n_x
2	356	2	296
3	182	3	155
4	96	3	85
5	45	4	47
6	23	6	25

When the number of consecutive rallies accounts for more than 90% of the overall rallies, it is reasonable to assume that in future forecasting models, gold can stop adding to positions when it has risen that many times in a row.
Let the probability that the number of consecutive rises exceeds the statistical count be greater than 90% set as u_1,then it can be expressed as u_1,

$$u_1 = \frac{m_x}{\sum_{x=2}^{5} m_x} \tag{9}$$

By the same token, we can get the probability of the number of consecutive decreases over 90%, which can be obtained by calculation, it is reasonable to think that when investing in gold,

ISSN: 2688-9323

adding, or reducing the position at most four times can ensure the maximum return without increasing the investment risk.

At this point, in the case of ensuring the number of positions, position reduction, if you want to maximize the benefits of investment, you need to statistics in the historical data to get four consecutive trading days of trading price increases and decreases, after we find the increase and decrease in the degree of the increase 9 quartile data and decrease 1 quartile data, respectively, recorded as $M_{0.9}$, , $M_{0.1}$.

And the 9^{th} percentile of continuous increase $M_{0.9}$ is 0.0447, 1st percentile of continuous decline $M_{0.1}$ is -0.0480.

Ideally, if we assume that each rise or fall is of the same magnitude, when the fall u_1 happens to be down $M_{0.1}$, then each fall can be added to the position, and the next day up $M_{0.5}$ can be gained.

Since commission needs to be considered for each position increase, then ideally, assuming that the amount of the first position increase is P_1 , the amount of the second, third and up to the nth position increase can be obtained.

The second position addition amount P_2 has the following relationship with P_1 ,

$$P_2(1-\alpha\%)M_{0.5} + P_1(1-\alpha\%)\frac{M_{0.1}}{u}M_{0.5} = 0 \tag{10}$$

Then the amount added after the nth position increase has the following relationship,

$$(1-\alpha\%)[P_n M_{0.5} + P_{n-1}\frac{M_{0.1}}{u_1}M_{0.5} + \cdots + P_1(\frac{M_{0.1}}{u_1})^{n-1}M_{0.5}] = 0 \tag{11}$$

According to $u_1 = 4$, $M_{0.1} = -0.0480$ above, the following table can be obtained.

Table 5. The relationship between number and amount

Number	1	2	3	4
Amount	P_1	$1.2P_1$	$2.88P_1$	$6.912P_1$
Amount	P_1	$1.1175P_1$	$2.4976P_1$	$5.5822P_1$

At this point, the regression model is introduced, and the number of positions added and the number of positions added are fitted to the curve, and the graph shows that the number of positions added is roughly exponential growth with the number of days the product price has fallen, so the curve is set as an exponential function.

Bringing in the data in the above table can be fitted to obtain the model of position addition amount

$$y = 0.1296P_1e^t \tag{12}$$

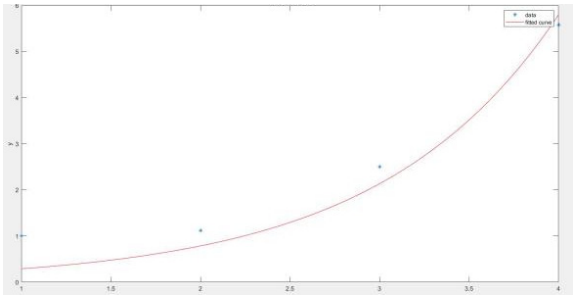

Figure 4. Gold Plus Position Management Model

ISSN: 2688-9323

Similarly, the position reduction model is obtained as,

$$y = 0.1062P_1e^t \tag{13}$$

In summary, the strategy of adding and subtracting positions for two financial products is obtained, then if the corresponding financial products are invested, the number of daily additions or subtractions can be found by the above fitting function, and the corresponding trading strategy is as described above.

Finally, to verify the feasibility of our proposed trading strategy, substituting the historical prices of gold and bitcoin over the past five years into the model yields the following return curve.

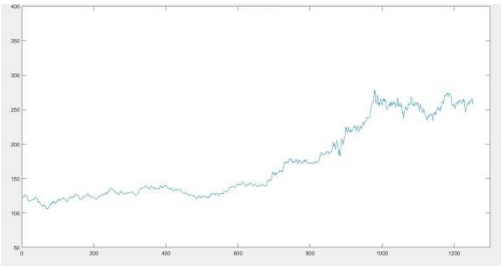

Figure 5. Return on earnings in one investment cycle

2.3. Sensitivity Analysis

Through the establishment of this trading model, it is easy to find that the model reacts well to the trading market of gold and bitcoin, and in the process of investment, because the increase and decrease of each trading day are different, the model gives different sizes of increase and decrease of positions, which can be well adapted to the law of market changes.

After the analysis of the return curve, the above proposed trading model is more stable and can give the best strategy to match the general environment of the short-term financial market in the trading strategy, avoiding certain riskiness.

For the transaction cost, it will affect the trading strategy we give, that is, when the transaction cost changes, the return obtained is not the same, but through image analysis we can conclude that when the transaction cost rises, the average return is decreasing, but the return curve keeps a steady rise, so it is reasonable to think that the model has a high sensitivity to the market and the trading scheme is reasonable and feasible.

Table 6. Relationship between transaction costs and benefits

Transaction Costs	0.2%	0.8%	1%	1.2%	2%
Average yield	450%	420%	400%	360%	335%
Earnings	4135.27	4196.58	4236.71	4123.87	4211.38

Note: The table still analyzes the return on a $1,000 trade over a five-year trading period, only the above relationship is obtained by changing the trading commission for gold.

3. Summary

By analyzing and mining the historical data, the initial amounts of the two financial product investment options were determined, after which the time series model was used to predict the exchange rate for the next day, and if the prediction was in an upward trend, the day was combined with the historical data to see how many days in a row the interval rose. According to the Apriori algorithm, four exponential functions are fitted, which can cope with different days of rise or fall respectively. The amount of the next day's position increase or decrease is

ISSN: 2688-9323

obtained by substituting the number of consecutive days.

The model is flexible in its response to the market and can make better feedback in the current market environment. By mining the historical data, we can make better trading plans with controlled riskiness in the market in the short term. And, the model can accurately analyze the relationship between the number of days of increase and the return, that is, the number of days of increase in the phase to determine whether to add or reduce positions, and the model in the process of running to add or reduce positions, the principal does not need to be taken out or deposited at once, which can effectively reduce the riskiness of the model.

It should be noted here that, through mathematical and statistical recognition, the probability of more than four days of continuous rise or fall is less than 10%, which is a small probability event, so this paper only calculates the first four days of continuous rise or fall of the number of positions added or reduced, if the continuous fall or rise of more than four days, no more trading until the financial product has turned over, and then brought back into the model calculation. This trading strategy generates a stable return based on risk avoidance, a higher return than buying a fund with fixed annual interest, and a high safety margin thus allowing flexibility to add and reduce positions on both sides of it.

References

[1] A fuzzy decision system for money investment in stock markets based on fuzzy candlesticks patter n recognition [J]. R. Naranjo, M. Santos. Expert Systems With Applications. 2019.

[2] Price manipulation in the Bitcoin ecosystem[J]. Neil Gandal, JT Hamrick, Tyler Moore, Tali Oberman. Journal of Monetary Economics. 2018.

[3] Least squares support vector machine for short-term prediction of meteorological time series[J] . A. Mellit, A. Massi Pavan, M. Benghanem. Theoretical and Applied Climatology. 2013 (1).

[4] Mining association rules between sets of items in large databases[J]. Rakesh Agrawal, Tomasz Imieliński, Arun Swami. ACM SIGMOD Record. 1993(2).

[5] Research and Application of Improved Apriori Algorithm Based on Matrix[J]. Yuan Liu, Yuan Sheng Lou. Applied Mechanics and Materials. 2014 (668).