*International Journal of Internet, Broadcasting and Communication Vol.14 No.4 140-147 (2022)*
*http://dx.doi.org/10.7236/IJIBC.2022.14.4.140*

**IJIBC 22-4-16**

# Smart contract research for data outlier detection and processing of ARIMA model

Youn-A Min

*Professor, Applied Software Engineering, Hanyang Cyber University, Korea*
*yah0612@hycu.ac.kr*

***Abstract***

*In this study, in order to efficiently detect data patterns and outliers in time series data, outlier detection processing is performed for each section based on a smart contract in the data preprocessing process, and parameters for the ARIMA model are determined by generating and reflecting the significance and outlier-related parameters of the data. It was created and applied to the modified arithmetic expression to lower the data abnormality. To evaluate the performance of this study, the normality of the data was compared and evaluated when the parameters of the general ARIMA model and the ARIMA model through this study were applied, and a performance improvement of more than 6% was confirmed.*

***Keywords:*** *Smart contract, AR, MA, ARIMA*

## 1. Introduction

According to 4th industrial revolution, the scope of application of smart devices and technologies is increasing, and thanks to this, cases of increasing the use and value of the overall industry such as factory automation using smart technology in industrial fields are increasing [1]. As such, as the application range of smart devices and technologies, which is increasingly applied in industrial fields, increases, the importance of data preprocessing for continuous operation and abnormal data detection for time series data and for maintaining normality of time series data is increasing [1,2].

In this study, as a method to increase the efficiency and normality of time series data operation applied in the industrial field, an accurate parameter measurement method of ARIMA by using a block chain technology-based smart contract for data preprocessing and detection and processing of outliers by section was studied.

<Related research>

Various industrial companies and factories predict future data through various types of data. In particular, when it is important to maintain a certain level of temperature and humidity, it is possible to infer whether the current value is normal or abnormal according to the pattern of existing data including time series data [1,2]. In this chapter, as a method of extracting and managing abnormal data by appropriately utilizing time series data, a smart contract based on block chain technology is used for data preprocessing to study a

Manuscript Received: September. 20, 2022 / Revised: September. 23, 2022 / Accepted: September. 25, 2022
Corresponding Author: yah0612@hycu.ac.kr
Tel: +82-2-2290-0872, Fax: +82-2-2290-0872
Professor, Applied Software Engineering, Hanyang Cyber University, Korea

method that can be utilized in the ARIMA model. In this chapter, the smart contract to be applied as an abnormal data filtering technology for data preprocessing and the ARIMA model, which is a representative model for processing abnormal time series data, and application cases of related technologies are reviewed.

## 2. Related work

### 2.1 ARIMA

Time series data means data prior to the present time, and time series analysis is a method of predicting a significant dependent variable through an independent variable [2], using past time series data as it is and within it. Look for patterns or trends. It is assumed that similar patterns will be reproduced in the future, and modeling and prediction are possible, so the utilization rate is high in various industrial sites [2,3].

Time series analysis has the characteristic of using time as an independent variable. In general time series analysis, it is assumed that the mean and variance are invariant regardless of time and that the covariance between two-time points is independent of other times in order to increase the stationariness of the data. It is also assumed that time series data probability characteristics are maintained up to the present time [3].

There are two general methods for time series data analysis, ARMA, and ARIMA [1-3].

The ARMA model considers AR (autoregressive) and MA (moving average), and AR is a model that allows the error term of previous observations to affect the observed data. The AR model has the formula X(t) = w*X(t-1) + b + u*e(t). The data for the current time point t is multiplied by the weight w of the previous time point data and a constant (b) is added [3]. e(t) refers to the current white noise and is a random value derived from a general normal distribution. AR models are often used for models that can infer continuous trends. The MR model has the formula X(t) = w*e(t-1) + b + u*e(t), and is a model suitable for a situation where trends change [3,4]. The equation of the MR model is calculated by reflecting the trend (w), not the current data (X), compared to the AR model. As such, since the ARMA method does not include an abnormal outlier removal process, the instability of time series data cannot be explained when there is a large number of abnormal data [2-4]. The ARIMA model, which is an improved ARMA model, includes the process of removing abnormal outliers in time series data [2-4].

ARIMA (Autoregressive Integrated Moving Average) is a model that considers both an autoregressive model and a moving average model, and uses the difference between observation data to explain non-stationary time series [3]. Usually, the Augmented Dlckey-Fuller Test (ADF Test) is performed to increase the normality of the time series data applied to the ARIMA model. The ADF test corresponds to a unit root test, and it is assumed that time series data can have a certain rule according to time. With respect to the random variable yt of the ADF test, the random variable at time t has a correlation with the random variable at time t-1 and t-2, but it means that errors may be included [1-3].

The parameters of ARIMA are usually p, q, d, where p, q represent the lag between AR and MA, and d means the difference [4]. In general, ACF and PACF functions are used to obtain p, q, and d, and recently, it is possible to output ACF and PACF and calculate appropriate parameters through a Python module [3-5]. The p value of a specific time point k can be obtained through ACF. The ACF uses the Cov function and the Var function for the time lag t and the specific time point k. PACF is a function used to exclude time lag t and observation values other than a specific time point k, and the measured values for q and d

operations can be calculated through the corr function [3-5]. With respect to AR(p), ACF has a form that exponentially decreases or disappears, and PACF has a form of cutting to zero after the parallax p. ACF and PACF calculated from MA(q) output the opposite situation to AR(p) [3].

Using ARIMA, it is also possible to observe white noise time series other than stationarity, which has a trend according to time, such as the effect of seasons.

### 2.2 Smart contract

Blockchain is a decentralized distributed ledger technology in which nodes shared in a network share, verify, and store data [6]. The blockchain can be divided into a public blockchain where all nodes that want a transaction can freely participate and verify, and a private blockchain where only authorized nodes can participate and verify a transaction [6-8]. A smart contract is a technology used in Ethereum, a type of public blockchain, and is an electronic contract function for use in applications of the blockchain Ethereum [7]. The smart contract writes the elements necessary for the contract in code so that the contract is automatically executed when data is input [7,12]. Smart contracts have tuning integrity and are executed via the Ethereum Virtual Machine (EVM). A fee, gas, may be generated when using a smart contract, but the fee may not be applied depending on the configuration of the blockchain platform [7,8].

The characteristics of a smart contract include observability that a smart contract should be able to observe each other's contract fulfillment potential or prove its performance, verifiability that it should be possible to verify when a contract is fulfilled or violated, and contract content privacy for information protection, and enforceability for enforcement and binding of contracts [7,10,11].

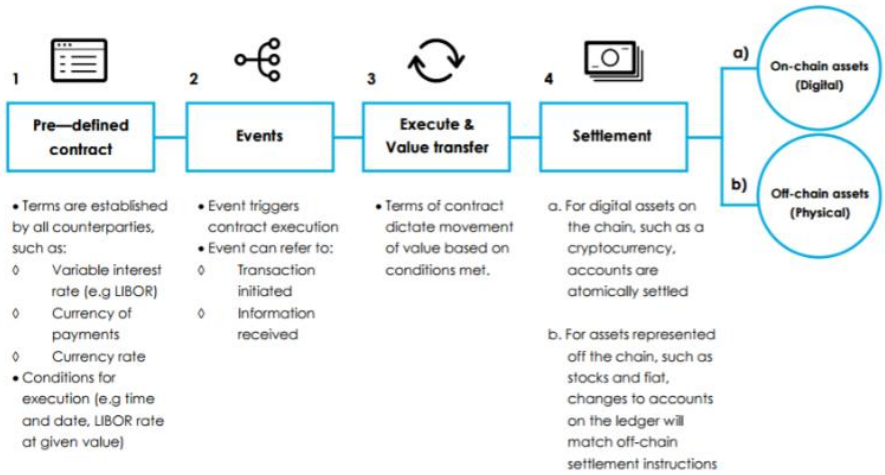



**Figure 1. Smart contract process [13]**

In this paper, the purpose of this study is to increase the accuracy of data pattern recognition and prediction by performing data preprocessing through smart contracts for each group divided into certain sections.

# 3. Smart contract research for data outlier detection and processing of ARIMA model

## 3.1 Modeling & Research method

As described above, a statistical function-based ARIMA model is used as a traditional method to utilize time series data. ARIMA is a parameter to obtain the average value of the MA and the lag of AR to measure the data error rate for a certain period of time. to calculate We also use the d parameter, which means difference. The typical values of p, q, and d of ARIMA have a distribution of p+q<2, p*q=0.

In this paper, a smart contract is used as a method to reduce the abnormal data tolerance rate that can occur when calculating the p, q, d parameters that depend on the existing function and to reduce the outliers for the abnormal data. Smart contracts have the feature of automatically executing transactions on input data, and through this processing, parameters can be calculated to efficiently process abnormal data among time series data. Let it affect the parameters p, q, and d to be applied.

The proposed process is as follows.

The process proposed in this study is as follows.

[Process] :

-Time series data to be used for evaluation are divided into groups with a time difference per section

-Give the significance (α) variable of the data group by section in the entire data

- Calculate the mean (A) and variance (B) for each group considering the significance (α) and outliers for the group

-Input the mean and variance of each group into the smart contract

-Generate parameter κ considering the number of outliers found, data characteristics, and significance of data groups (> 0, <1)

-Applying the importance (α) and the parameter (κ) indicating the interval outlier through the modified formula when measuring the ARIMA parameter

**Figure 2. Proposed process**

The modified formula for ARIMA parameter measurement during the process of Figure3 is shown in Figure 3.

- $r\_p = \frac{Cov(y_t, y_{t+k})}{Var(y_t)} \times (1-k) * \alpha$
- $r\_q = \frac{corr(y_t, y_{t-k} \mid y_{t-1}, y_{t-2}, \ldots, y_{t-k+1})}{corr(y_t, y_{t-k} \mid y_{t-1}, y_{t-2}, \ldots, y_{t-k+1})} \times k$
- $p = p / r\_p$
- $q = q / r\_q$

**Figure 3. Proposed formula for ARIMA**

In Figure 3, it is possible to calculate in consideration of the significance of the data group and the ratio of outliers.

### 3.2 Experiment and performance evaluation

The data structure for the performance evaluation of the contents of this study is organized as shown in Figure 4. 8G RAM and Linux virtual environment are used as computing specifications for the experiment.

- data: Humidity data that has been output normally for a certain period of company A and organized into time series data
- Applying white noise to normal data: Intentionally applying white noise to data less than 3% of normal data
- Experimental method:
  - Method 1 ) Traditional ARIMA parameter measurement and ACF measurement
  - Method 2) ACF measurement after smart contract-based data preprocessing and ARIMA parameter measurement

**Figure 4. Experiment environment**

In order to determine p, q, and d in traditional ARIMA, ACF (Autocorrelation function) Plot, which measures the relationship between observation data according to lag, and PACF ( Partial autocorrelation function) Plot is used. In the case of time series data with AR characteristics, ACF decreases slowly and PACF decreases rapidly. In the case of data with MA characteristics, on the contrary, PACF decreases slowly and ACF decreases rapidly. These lags are used as parameters p and q, and the data difference and the number of differences can be obtained. For the performance evaluation of the contents, statsmodels, a python package for ACF and PACF calculations, was used, and part of the code for ACF and PACF in group A is shown in Figure 5.

```
import matplotlib.pyplot as plt
import pandas as pd
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
series = pd.read_csv('score_a.csv', header=0, index_col=0, squeeze=True)
series.plot()
plot_acf(series) #1)
plot_pacf(series) #2)
...
plot_acf(diff_t) #3)
plot_pacf(diff_t) #4)
```

**Figure 5. Part of code about acf / pacf**

The process for constructing the ARIMA model that has undergone data preprocessing through smart contracts will be described. When applying the ARIMA model, it is important to remove outliers for the time series data to increase the validity. In this study, a smart contract based on blockchain technology was applied for efficient detection and processing of outliers by section.

The input time series data is divided into groups and grouped, and the characteristic value for the group is processed as an input value to the smart contract.

Figure 6 is part of the smart contract code for the process

```
// Set the weight according to the significance of the data for each section
Weight.Sent().watch({}, "", function(error, result) {
if (!error) {
console.log("...
}
})
// gkatn that changes the parameter value when abnormal data for each time series section appears above the threshold
function ChangePara(
uint _biddingTime,
address _beneficiary
) public {
...
}
‘’‘
```

**Figure 6. Part of code about smart contract**

According to the ratio of outliers in the time-series data grouped by period, a parameter (κ) that can indicate data normality is generated and assigned, and the significance of the data group is considered in the process.

When generating parameters, when the mean and variance of a certain section (y) have an error rate of 30% or more with y-1, y-2, y+1, y+2, and the frequency of outliers compared to normal values is 60% or more The interval is treated as the average value of [y-1, y-2, y+1,y+2], and when the frequency of outliers is over 90%, sequentially to the abnormal group so that it cannot be applied to separation and prediction for the data group include as

The following is a part of Python code modeling by applying ARIMA.

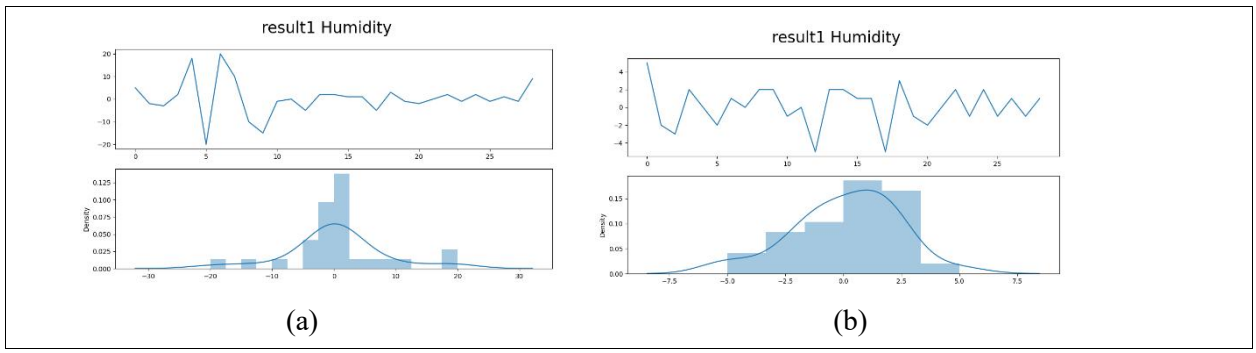
```
    from statemodels.tsa,arima_model. importARIMA
    import statsmodels.api as sm
    train=ARIMA(score_a_test['score'],order=(1,2,0))
    train_fit=model.fit(trend='c', full_output=True, disp=True)
    train_fit.summary())
    ...
train['feature_name'].diff().dropna()
```

**Figure 7. Part of code about ARIMA**

As shown in the parameter generation formula in Figure 3, it is possible to determine whether to drop even in the case of abnormal data occurrence according to data importance, and it is possible to check the exact parameters of the data included in the section according to the degree of outlier parameters for each section.

Since the characteristics of the data indicate daily learning satisfaction, the correlation can be compared according to the ACF. Figure 8 shows ACF and PACF through ARIMA with a lag of 10 and automatically obtained parameters.

As shown in the figure 8-(a), the correlation is deep, but there are several abnormalities in the data.




(a) (b)

**Figure 8. Performance evaluation**

Figure 8=(b) shows ARIMA's ACF and PACF applied after obtaining parameters by applying the parameters according to the data abnormality rate by applying the smart contract proposed in this study with a lag of 10. And in the deep correlation is the same, but it can be seen that the normality of the data is improved.

Table 1 shows the accuracy and error rates of the traditional ARIMA method (test1) and the proposed ARIMA application (test_2) based on the code in Figure 8.

**Table 1. Accuracy / Error rate about evaluation**

| Var | Accuracy (%) | Error rate (%) |
|---|---|---|
| test_1 | 92.5 | 7.5% |
| test_2 | 93.1 | 6.9% |

As the result of Table 1 shows, as a result of applying data preprocessing through smart contract, it can be confirmed that the accuracy rate is increased by 8.5~9.5% compared to the existing method.

## 4. Discussion

In this paper, for accurate prediction of time series data, outlier detection and missing value processing were studied by smart contract code before the existing ARIMA method. For time-series data divided into sections, parameters representing stationary and non-stationary data were derived through smart contracts, and the ARIMA parameter calculation method was modified according to the importance and non-stationary nature of the data to increase data normality. As a result of this study, it can be confirmed that the accuracy rate of data prediction is increased by more than 8.5% compared to the existing method. As a result of this study, the accuracy of data prediction is increased, but the economic feasibility of operation due to smart contract operation may decrease. In the future, it is planned to supplement the scalability of the study by diversifying the characteristics of the dataset and the period of collected data as well as methods to increase economic feasibility when processing data on abnormal data groups.

## References

[1] Hyuncheol Jeong et al., "Development of ARIMA-based Forecasting Algorithms using Meteorological Indices for Seasonal Peak Load", The Transactions of the Korean Institute of Electrical Engineers v.67 no.10, 2018, pp.1257 - 1264 ,2018.
[2] Siyeon Kim et al., "Weekly Maximum Electric Load Forecasting for 104 Weeks by Seasonal ARIMA Model", Journal of the Korean Institute of IIIuminating and Electrical Installation, Vol.28, No.1, pp.50-56, 2014.
[3] Smart-contract, Available: https://youteam.io/blog/5-essential-steps-for-successful-smart-contract-development/
[4] https://www.lgcns.com/blog/cns-tech/30841/
[5] Zheng, Xiuyan et al., "Stock Trend Prediction Based on ARIMA-LightGBM Hybrid Model",2022 3rd Information Communication Technologies Conference
[6] Park Geun-chae, Baek Jun-geol, "Time Series Prediction using ARIMA and DBNs with MODWT", Journal of Korean institute of industrial engineers v.43 no.6, pp. 474 - 481
[7] Grillenzoni, Carlo, "ARIMA Processes With ARIMA Parameters",Journal of business & economic statistics: a publication of the American Statistical Association v.11 no.2 ,pp. 235 - 250 , 1993 ,
[8] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system”, White Paper, 2008, Available: Available: https://bitcoin.org/bitcoin.pdf.
[9] V. Buterin, Chain Interoperability, Available: https://www.bubifans.com/ueditor/php/upload/file/20181015/1539602892605747.pdf.
[10] Lim Jong-cheol, Yoo Hyun-kyung et al., "Blockchain and Consensus Algorithm", Electronics and telecommunications trends v.33 no.1, pp.45 - 56 ,2018.
[11] Shin Eun-seop, "Results of preliminary feasibility study for long-term technology development project of block chain", Information sciences v.519, pp.348-362, 2019.
[12] L. Luu, PeaceRelay: Connecting the many Ethereum Blockchains, 2017, Available: https://medium.com/@loiluu/peacerelay-connecting-the-many-ethereum-blockchains22605c300ad3.
[13] Sundeok Yoo, " A Study on Consensus Algorithm based on Blockchain", The journal of the institute of internet, broadcasting and communication, JIIBC Vol.19, No.3, pp.25 - 32, 2019.
[14] P. Ruan, G. Chen, T. T. A. Dinh, Q. Lin, B. C. Ooi and M. Zhang, “Fine-Grained Secure and Efficient Data Provenance on Blockchain Systems”, PVLDB, Vol. 12, No. 9, pp. 975-988, 2019.
[15] smart contract processing, 2020, Available: https://www.lgcns.com/blog/cns-tech/30841/