CCC Publications

A financial time series data mining method with different time granularity based on trend Division

Haining Yang, Xuedong Gao, Lei Han, Wei Cui

Haining Yang*

Department of Management Science and Engineering,
School of Economics and Management,
University of Science and Technology Beijing, Beijing 100083, China
*Corresponding author: yanghaining@apiins.com

Xuedong Gao, Lei Han

Department of Management Science and Engineering,
School of Economics and Management,
University of Science and Technology Beijing, Beijing 100083, China
yanghaining@apiins.com, gaoxuedong@manage.ustb.edu.cn

Wei Cui

School of Economics and Management,
China University of Geosciences (Bei-jing),
Beijing 100083, China
cuiw@cugb.edu.cn

Abstract

Stock research is an important field of Finance and time series research. Stock data research is also a typical financial time series problem. In the research of financial time series, there are many methods, such as model building, data mining, heuristic algorithm, machine learning, deep learning, and so on. VAR, ARIMA and other methods are widely used in practice. ARIMA and its combination methods have good processing effect on small data sets, but there are over fitting problems, which are difficult to process large data sets and data with different time granularity. At present, this paper takes the decision table transformation method of financial time series data as the research object, and puts forward the trend division method of financial time series based on different time granularity through the trend division of financial time series. On this basis, it puts forward the trend extreme point extraction method, and constructs the stock time series decision table according to the extreme point information and combined with the stock technical indicators, The decision table is verified by support vector machine based on the decision table. The research shows that the trend division method under different time granularity can transform the extreme point information into a decision table, which will not produce over fitting problem in practical application. It is an effective time series processing method, and provides a new research method for the future time series research with different granularity.

Keywords: Time series, Trend division, Time granularity, Arima, Support vector machine.

1   Introduction

Time series analysis is an important research field of traditional statistics. After decades of development, it has laid its own theoretical foundation. The research object of traditional time series analysis is generally dynamic data with randomness, and the research method focuses on the construction of global model. At present, the research of time series mainly focuses on econometric methods and data mining methods. However, in the actual analysis of time series data, it is often necessary to analyze the local characteristics of time series, such as finding the frequent patterns, comparing different sequences or whether the same sequence has similarity in different time periods, etc.

Time series records the state values of observation targets at each time point to form complex high-dimensional data. It can be divided into two categories. One is simple time series, which is stable and linear; The other is complex time series, which is non-stationary, nonlinear and multidimensional. The process of analyzing the current observation value of time series and inferring the value of the observed object at a future time node is called time series prediction.

Time characteristic is an important feature of time series data. The time characteristic of data is not only an index but also a division basis, and it has the characteristics of independent variable and dependent variable. Time series data also has the characteristics of multi time granularity. Taking stock data as an example, there are data with different granularity, such as 1 minute, 5 minutes, 15 minutes, 30 minutes, 1 hour, day, week, month, quarter, year and so on. There is some uncertain relationship between various granularity, which also provides a new research direction for time series prediction.

Stock prices, trading volume indicators and many new indicators derived from them can better reflect the internal relationship of stock prices. Making full use of these indicator data often has certain guiding significance for the decision-making of individual stock trends, choices and trading time points. Based on this, it is necessary to further study the multi granularity time series forecasting method based on trend division and combined with a variety of indicators, establish a reasonable and effective stock price forecasting model, enrich and develop the theory and method of stock time series research, and provide a reliable basis for investors' decision-making.

To sum up, this paper will process time series with different time granularity according to the trend division of time series, extract extreme points from general time series, build decision table, and mine stock time series according to the decision table, hoping to provide a new method for processing financial time series.

2   Related work

2.1   Research on univariate time series

In the field of econometrics, univariate time series modeling has become a very important research field in economic management research [1, 2, 3]. Time series analysis is a statistical method to reveal the dynamic structure and regular characteristics of the system according to the fluctuation of dynamic data. The core idea is: according to the limited length of time series data observed by the system, establish the corresponding data model to reflect the dynamic dependencies contained in the time series, and thus predict the future trend of the system. In the development of time series model, an important feature is the assumption of statistical equilibrium relationship, one of which is the assumption of stationarity. Usually, a stationary time series can be effectively described by its mean, variance and autocorrelation function. In general, the main methods of data research are model construction [4, 5] and data analysis. In the next part, we will review different methods.

Financial time series is an important research field of time series, which has many data patterns. The research on the trend and price of stocks is an important research hotspot of financial time series. At present, the research of stock time series focuses more on fitting the series itself and making time series prediction based on it. Under this research paradigm, the prediction is often effective in a certain time range, but the generalization ability is poor. For example, the differential integrated moving average autoregressive model (ARIMA) can deal with non-stationary time series, but because the processing process of time series' stationarity masks the complex and diverse characteristics of

fluctuation patterns in the series, it is also unable to make a more accurate prediction for complex and changeable time series. The discrete Fourier transform (DFT) algorithm is simple to calculate, and can concentrate most of the energy of the time series signal into a few coefficients. The time complexity is also relatively low, but the algorithm discards the high-frequency component of the signal in the process of data interception, smoothes the local maximum and minimum of the signal, and causes the omission of information, and it is not suitable for non-stationary sequences [6]. Discrete wavelet transform (DWT) can consider the local characteristics of time domain and frequency domain at the same time. On the other hand, under the same data compression, DWT can better retain local details and provide more accurate approximate results than discrete Fourier transform. Discrete wavelet transform does not reduce the relative image error, nor does it improve the accuracy of similarity query. From the experimental results, the effect of discrete wavelet transform and discrete Fourier transform is not very different [7].

In the 1970s, ARIMA (autoregressive integrated moving average model) was proposed by G.E.P.Box, an American statistician, and G.M.Jenkins, a British statistician [8]. The time series prediction method using this model is also called Box-Jenkins method. ARIMA (P, D, q) model includes autoregressive (AR) process, moving average (MA) process, autoregressive moving average (ARMA) process and ARIMA process. Where p is the autoregressive term, D is the number of differences when the time series is processed smoothly, and Q is the moving average term.

The stability of ARMA (P, q) process depends on the autoregressive part, while its reversibility depends on the moving average part. The nonstationary of time series is various, but the economic time series often has the characteristics of linear homogeneous nonstationary. If a random process contains D unit roots, the random process can be transformed into a stationary autoregressive moving average process after D times of difference, and the random process is a single integer autoregressive moving average process, which is recorded as ARIMA (P, D, q). However, these time series analysis methods require linear models. In order to overcome this defect, nonlinear time series models have become a research hotspot.

Granger [9] The bilinear time series model is proposed. By using the residual information generated in the process of time series prediction, the model has the characteristics of concise structure, few parameters and strong adaptability in practical application, which makes the model have wide application value in nonlinear time series analysis. With the development of interdisciplinary research, neural network, wavelet decomposition, Fourier transform and fractal theory and methods are used for data mining and prediction of time series.

From the existing literature and the mathematical expression of the above time series, it is found that:

1. The premise of time series analysis is to assume that the past of variables will extend to the future, that is, the future of the system will not suddenly change by leaps and bounds, but a gradual process.

2. In time series analysis, we should first test the stability of time series, select the appropriate time series model after appropriate data processing (such as difference) and passing the test, and estimate the parameters in the model.

The prediction of stock prices and trends mainly revolves around several aspects. One is technical indicators and feature engineering. Hossain, e., et al. Used the technical analysis of expert system based on belief rules (BRBES), combined with the concept of brin belt, to predict the stock prices in the next five days, and studied the effects of different deep learning methods [10]; Ji, g., et al. Studied the application of Feature Engineering in stock prediction, proposed improved technical indicators based on wavelet denoising and a new two-stage adaptive feature selection method, and adopted random forest model as the stock prediction model [11]; Li, G. Z., et al. Proposed a multi index feature selection method for stock price prediction based on Pearson correlation coefficient (PCC) and generalized learning system (BLS) [12].

The methods of stock forecasting mainly focus on deep learning and fusion models. Deng, C. R., et al. Developed a hybrid stock price index prediction modeling framework using long-term and short-term memory (LSTM) and multivariate empirical mode decomposition (MEMD), which can

capture the intrinsic characteristics of the complex dynamics of the stock price index time series [13]; Gao, R. Z., et al. Proposed a deep learning method combined with genetic algorithm to predict the target stock market index [14]; Gao, Z., et al. Proposed a prediction algorithm integrating multiple support vector regression (SVR) models, and used reasonable weight to combine the prediction results of multiple models to improve the accuracy of the model [15]; Gupta, U., et al. In order to overcome the problem of overfitting, a new data enhancement method was proposed in the StockNet model based on Gru [16]; He, Q. Q., et al proposed a new case-based deep transfer learning model with attention mechanism [17]; Kanwal, A., et al. Proposed a prediction model based on hybrid deep learning (DL), which combines deep neural network, short-term memory and one-dimensional convolutional neural network (CNN) [18]; Kumar, R., et al. Proposed a three-stage fusion model to process time series data and improve the accuracy of stock market prediction [19]; Li, R. R., et al. Proposed a multi-scale modeling strategy based on machine learning methods and econometric models [20].

There are some new methods applied to stock time series prediction. Liang, M. X., et al. Used sequence morphology mining to obtain Candlestick morphology from multidimensional Candlestick data, calculated the correlation between different morphology and corresponding future trends, and proposed a new sequence similarity, matching different candle graph sequences with existing patterns [21]; Wang, C. J., et al. Use the deep learning framework transformer to predict the stock market index. Through the encoder decoder architecture and multi focus mechanism, research shows that transformer can better describe the basic rules of stock market dynamics [22].

2.2 Literature review and analysis

Through the above literature review, we find that time series has become a more and more popular research field, especially the rise of deep learning technology, which partially solves the problem of time series' interconnection in time scale, and time series research has become an independent research direction. However, the traditional time series research attempts to use a single or multiple models to explain or characterize the time series, and make time series prediction accordingly. Under this research paradigm, the prediction is often effective within a certain time range, but the generalization ability is poor (only effective for the trained data, and poor for the new data), which cannot meet the requirements of the era of big data. Machine learning methods need to rely on artificial feature selection. Although the deep learning algorithm has achieved good results, it is still in a black box state with weak interpretability. Moreover, the deep learning algorithm is also prone to butterfly effect. A little change can have a significant impact on the results of the deep learning model, it is also easy for decision makers to make wrong conclusions. Some scholars also began to explore the combination of deep learning algorithm and traditional methods to achieve better results.

3 Methodology

3.1 Comparative analysis of multiple time granularity prediction based on ARIMA model

As one of the traditional time series prediction models, ARIMA model has the following general steps:

1. Stability test. The ADF unit root test is used to check whether the sequence is stable. If the sequence is unstable, the difference operation is performed.

2. White noise inspection. Check whether the sequence is random.

3. Model order determination. ACF and PACF functions are used to determine the order of the model and determine the final ARIMA model.

4. Model prediction. The ARIMA model is used to predict the sequence data.

There are many evaluation indexes for the performance of the model. In this paper, the mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) and the coefficient of determination R2 are selected to evaluate the performance of the model.

This study selects the Growth Enterprise Market (GEM) index (399006) 1-minute interval, 5-minute interval, 15-minute interval, 30-minute interval, 60-minute interval and daily interval data, including the date from August 19, 2019 to August 19, 2022. Specific data statistics are shown in Table 1:

Table 1: Basic data

Data introduction	
data sources	Wind database
Number of 1-minute interval data	176176 pieces
Number of 5-minute interval data	34944 pieces
Number of 15-minute interval data	11648 pieces
Number of 30-minute interval data	5816 pieces
Number of 60-minute interval data	2912 pieces
Number of daily interval data	728 pieces

In this paper, the model is built for different granularity data sets, and the performance of the model is statistically analyzed.

(1) ARIMA model of daily data

This paper uses Python language to write ARIMA model, and takes the last 10 data of the data set as the test set to judge the generalization ability of the model. First, the stability test of the sequence is performed. Table 2 shows the results of ADF unit root test:

Table 2: ADF unit root test of daily line data

Differential order	T	P	critical value		
			1%	5%	10%
1	-26.887	0	-3.439	-2.865	-2.569
2	-16.557	2.256	-3.439	-2.865	-2.568

The white noise test for the stationary sequence after the first-order difference shows that the p value is 2.4656815e-05, which is significantly less than 0.1, proving that the sequence is non pure random sequence, which can be used for model construction. The sequence diagram after 1st order difference is shown in Figure 1:
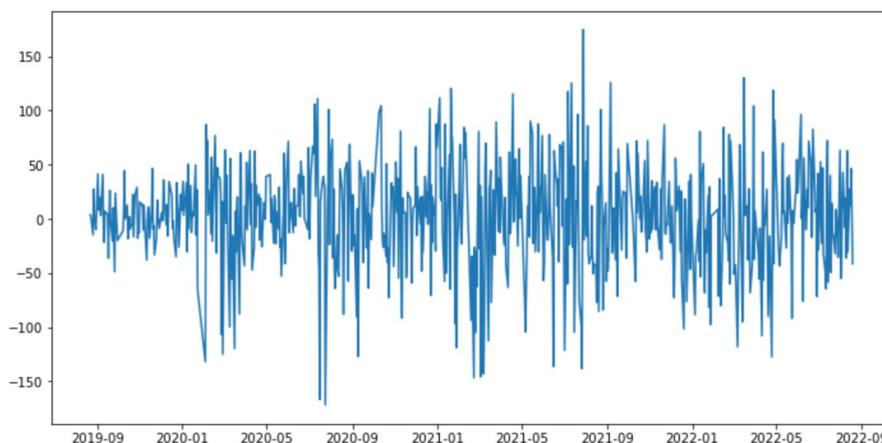

Figure 1: 1st order differential sequence diagram of daily interval data

According to the data after the first-order difference, the ACF and PACF functions are used to determine the order of the model. The ACF and PACF diagrams are shown in Figure 2:
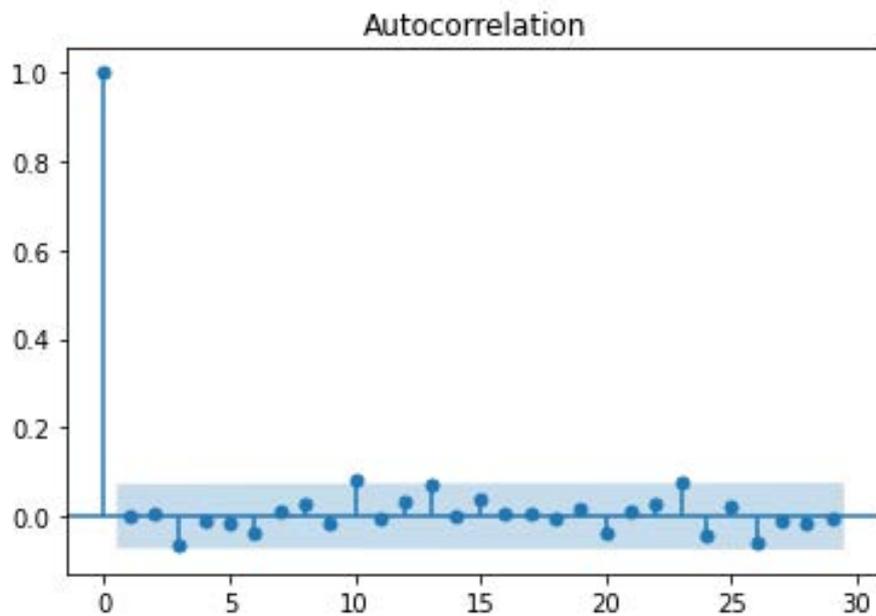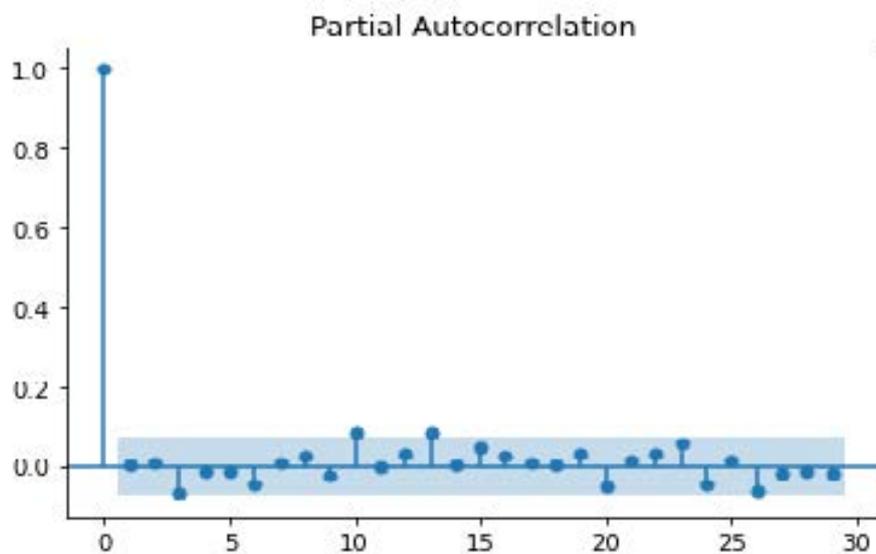
Figure 2: Daily interval data ACF

Figure 3: Daily interval data PACF

Observing the ACF chart and PACF chart in Figure 3, using the BIC criterion with objective criteria to determine the order of the model, it is found that when $p = q = 0$, the amount of BIC information achieves the minimum value, and the model achieves the optimal fitting effect, that is, the fitting model of the original series is Arima (0,1,0), and it can also be seen that the stock time series meets the random walk model, That is, the position of the current time is only related to the position of the previous time.

The ARIMA model obtained is used for fitting and prediction, and the fitting results are shown in Figure 4:
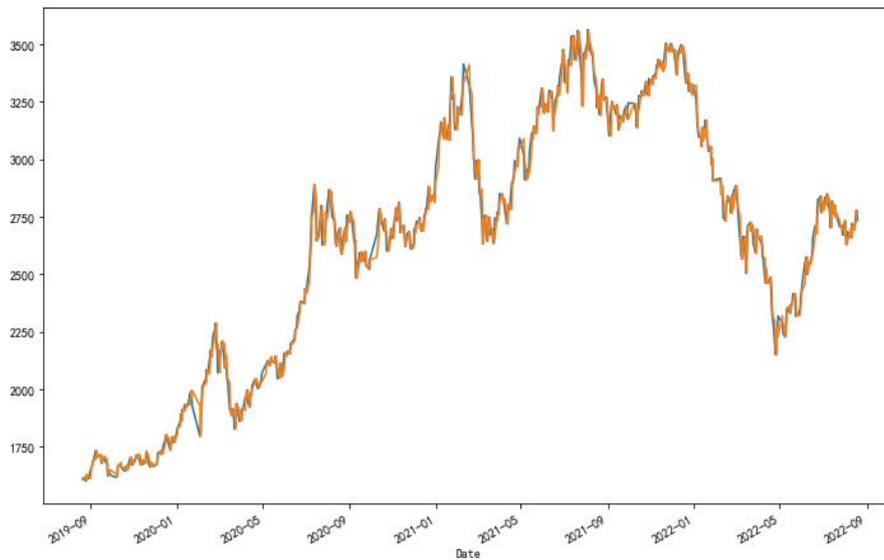
Figure 4: Fitting results of daily line model

The prediction of the model on the test set is shown in Table 3:

Table 3: Forecast value of daily interval model

time	True value	predicted value	Absolute value of fitting error
2022/8/8	2675.686	2682.774	7.0876
2022/8/9	2694.8	2674.883	19.9173
2022/8/10	2658.585	2693.935	35.3507
2022/8/11	2721.493	2657.838	63.6549
2022/8/12	2690.831	2720.541	29.7095
2022/8/15	2718.593	2689.979	28.6138
2022/8/16	2731.393	2717.651	13.7426
2022/8/17	2777.91	2730.409	47.5004
2022/8/18	2775.821	2776.774	0.9534
2022/8/19	2734.218	2774.692	40.4743

Model performance indicators are shown in Table 4:

Table 4: Performance indicators of daily interval model

index Dataset	MAE	MSE	RMSE	MAPE
Model training set	37.8757	3270.24	57.186	0.0145
Model training set	33.962	1153.419	33.962	0.0105

Table 4 shows that the performance of the model on the training set is excellent and can basically meet the requirements, but the prediction effect on the test set is not particularly ideal, and the prediction error is large, reflecting the poor generalization ability of the model.

Since the modeling process of other granularity data is consistent with the daily data, this paper will not repeat the research process, but the performance index statistics of ARIMA models with different time series granularity are shown in Table 5:

Table 5: Performance indicators of ARIMA model with different granularity

index Dataset	Mae	MSE	RMSE	MAPE
1Minute interval model training set	-	-	-	-
1Minute interval model training set	-	-	-	-
5Minute interval model training set	4.1735	79.4054	8.9109	0.0015
5Minute interval model test set	3.0347	12.84771	3.5843	0.0011
15Minute interval model training set	7.5064	171.1889	13.0839	0.0028
15Minute interval model test set	6.956	62.6869	7.9175	0.0025
30Minute interval model training set	11.0193	354.5312	18.8289	0.0041
30Minute interval model test set	7.342	83.2282	9.1229	0.0026
60Minute interval model training set	16.3431	749.5902	27.3786	0.0062
60Minute interval model test set	9.1569	150.6016	12.2719	0.0033

Note: - indicates the amount of data that the computer cannot process.

According to the performance index statistics of ARIMA models with different granularity, with the decrease of granularity, the amount of data increases significantly, and the MAE, MSE and RMSE of the training set model all show a downward trend. However, compared with the training set, the prediction effect of the model on the test set is poor, reflecting the over fitting phenomenon of the model, and with the increase of the amount of data, the over fitting phenomenon is becoming more and more obvious. This also reflects that the traditional time series research attempts to use a single or multiple models to explain or characterize the time series, and make time series prediction accordingly. Under this research paradigm, the prediction is often effective in a certain time range, but the generalization ability is poor, and it is difficult to deal with multi granularity time series effectively.

3.2 Establishment of multi granularity time series decision table based on trend partition

In 2011, Preis et al [23] A time renormalization method is proposed to analyze the switching dynamics of the German futures index (FDAX). Its research shows that the switch point in the German futures market index tends to have large volatility, trading volume and intensive trading time interval.

First, at the price $P(t)$ of time t, if we want to judge whether $P(t)$ is a turning point, we must compare the prices of a window size (S) before and after $P(t)$, that is, $P(t-S)$ and $P(t+S)$. If there is no price higher than any price $P(t)$ in the range of $P(t-S)$ to $P(t+S)$, then any price $P(t)$ is defined as the highest price $P_{\max}(S)$. In the same way, from $P(t-S)$ to $P(t+S)$, if no price is lower than any price $P(t)$, then any price $P(t)$ is defined as the lowest price $P_{\min}(S)$, and the turning point we mentioned earlier is the maximum price $P_{\max}(S)$ and the minimum price $P_{\min}(S)$ defined above. As shown in the figure.

Schematic diagram of price change over time: in the stock price trend chart, select a window size (S), take s as the range, and judge whether $P(t)$ is the minimum or maximum value within this range from $P(t-S)$ to $P(t+S)$. It can be seen from the figure that in $P_{\max}(S)$ and $P_{\min}(S)$, the blue line is $P_{\max}(S)$, and the red line is $P_{\min}(S)$.
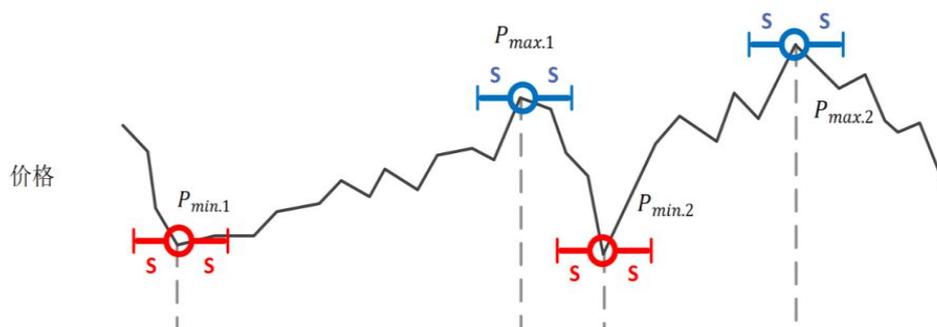
Figure 5: Price and trend turning point

Next, further discuss the concept of trend from the turning point, as shown in Fig 2 $P_{\min}.1(S)$ to $P_{\min}.1(S)$ The price fluctuation between them showed an upward trend, followed by a downward trend to $P_{\min}.2(S)$. In this section, we regard the price change of the Taiwan index period as consisting of several upward and downward trends. Within a certain range, each trend contains this extreme point, and this local extreme point forms the reverse trend of the time series. Identifying these extreme points has important research value.

3.3 Extraction of time series trend extreme points based on time series value recursion

Time series data is data that grows with time. The longer the data exists, the larger the amount of data. It is very important to find data points that can represent characteristics and have reference value from these data when analyzing time series data. The reason is that the data can be compressed and less data can show the trend of the overall huge amount of data. The trend division of time series should first consider the characteristics of time series. Taking the stock price of financial time series as an example, the general stock information is expressed in the form of one minute, five minutes, fifteen minutes, thirty minutes, days, weeks, months and years. In different time granularity, there are often great differences in the division of trends.

The original time series trend division method generally revolves around several key points, such as segmentation range, sliding window, segmentation point and threshold determination. These contents make the trend division of time series full of uncertainty. Although the results can be continuously optimized through multiple experiments, a structured trend division method has not been formed, and its stability is difficult to ensure in the process of use. However, in the subsequent links such as data prediction, the over fitting phenomenon will be caused due to different parameters.

Dow theory points out that the main trend, medium-term trend and short-term trend are the three main manifestations of stock price changes, and the newly formed bottom and top prices are the main basis for judging the formation of the trend. The original time series trend division method generally revolves around several key points, such as segmentation range, sliding window, segmentation point and threshold determination. These contents make the trend division of time series full of uncertainty. Although the results can be continuously optimized through multiple experiments, a structured trend division method has not been formed, and its stability is difficult to ensure in the process of use. However, in the subsequent links such as data prediction, the over fitting phenomenon will be caused due to different parameters.

According to the characteristics of time series, this paper proposes a new time series trend division method. The main idea is that a trend must have extreme points, and a local maximum extreme point and two local minimum extreme points before and after constitute an upward (or downward) trend. Therefore, in the process of trend recognition, it is particularly important to find the extreme points in the time series. This paper uses Python to write programs to find the extreme points of time series, and then identify the rising and falling trends in time series. The specific algorithm is shown in Table 6:

Table 6: Extraction of time series trend extreme points based on time series value recursion

Algorithm:

Input: original time series data $s = \{s_1, s_2, ..., s_n\}$, step M
Output: index k of maximum point and index H of minimum point of time series s.

(1)   For
       $I = \{1, 2, 3, ..., n\}$, do
(2)   If $s_1 > s_1 + m$,
(3) $K \leftarrow 1$
(4)   Else if $s_1 < s_1 + m$,
(5) $H \leftarrow 1$
(6)   Else if $s_i - m < s_i < s_i + m$,
(7)   Break
(8)   Else if $s_i - m < s_i > s_i + m$,
(9) $K \leftarrow I$
(10) Else $s_i - m > s_i < s_i + m$,
(11) $H \leftarrow I$
(12) Organize index K, H
(13) End

In order to verify the feasibility of the above algorithm, a group of time series data is randomly selected to extract the extreme points. Taking the time series of [1, 6, 30, 20, 15, 7, 13, 9, 2, 6, 0, -7, -15, -6, 4, 13, 8, 15, 8, 12, 3,0, 20, 7, 3, 9] as an example, using the method proposed in this paper, the step size is 1, and the extreme points are identified. The visualization results are shown in Figure 6.
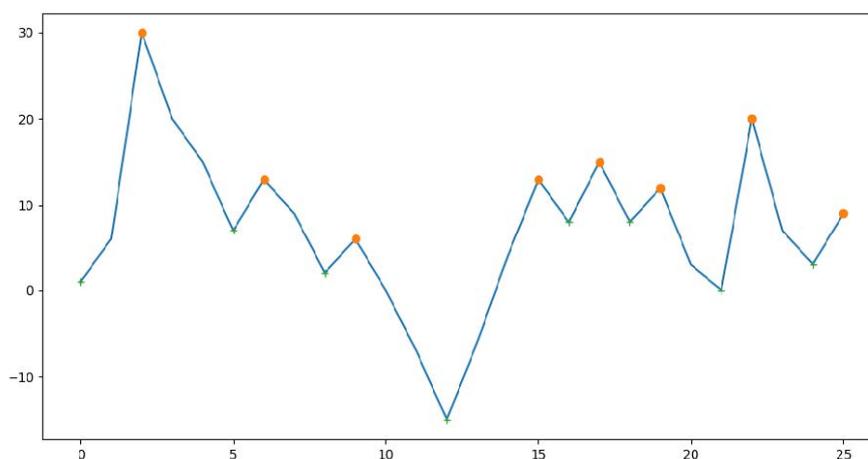

Figure 6: Visualization of extreme point extraction of random time series

The extreme points extracted are: maximum points: [30,13,6,13,15,12,20,9]; Minimum point: [1,7,2, -15,8,8,0,3]. The results are consistent with the expected results, which shows that the proposed method is effective.

The decision table, also known as the judgment table, is suitable for describing the situation where there are too many judgment conditions and the conditions are combined to form a variety of decision schemes. The decision table is composed of state variables (condition items) and decision variables (action items), and each state variable and decision variable together constitute a rule. The decision table can list complex problems according to various possible situations. Using the decision table, the state variables can be understood as inputs and the decision variables as outputs. This is more similar to the idea of machine learning in data mining, so decision table can be widely used in classification, clustering, association rule mining and other fields.

1. Determination of decision variables

The above algorithm is used to process the time series data set. When the step size is set to 1 at runtime, the local maximum and local minimum of trends with different time granularity can be obtained. The maximum point is represented by "o", and the minimum point is represented by "+". The idea of two classification is adopted, and the local maximum and local minimum are used as the decision variables of the information table. At the same time, for the stock time series data set, the only intuitive information we can get is the real-time stock price and trading volume. If only the price and trading volume are included, the information in the decision table is very little, which is difficult to meet the requirements of data processing and effectively predict the decision variables. Therefore, we need to further enrich the information table, Select more indicators related to decision variables to form the information table.

2. Determination of factors in decision table

In the current stock market research, some people have studied it from the perspective of enterprise management and finance, forming a fundamental analysis method, while others have studied it from the perspective of stock price and form construction, forming a technical analysis method. Stock technical analysis is based on derivative data as the research object, for example, by investigating the charts of data such as stock price and trading volume, and predicting the future trend according to the stock history and current performance. Technical indicators provide data support for investors who analyze stocks from the intraday. Therefore, adding the existing empirical indicators of stock technical analysis to the decision table will greatly increase the content of the decision table and enable it to use the decision table to carry out data mining.

In technical analysis, the most important factors include price, quantity, time and space.

- The most basic element of technical analysis is price, which is also one of the key elements of the stock market; The intrinsic value of stock is subject to change, and its fluctuation is jointly affected by the operating conditions of the stock issuer and the behavior of the stock market. When the stock issuer has a significant positive or negative effect, the stock chips will have a corresponding effect, which will be directly reflected by the rise and fall of the stock price.

- Another important element of technical analysis is trading volume. Trading volume is an important factor affecting the formation of stock prices. The standard for confirming the price form is trading volume. In the stock market, only buying and selling can lead to price fluctuations. When the stock price is low and there are a large number of purchases, it is generally believed that the stock price will rise. When the stock price is relatively high and there are a large number of sales, the stock price will generally fall. Therefore, investors generally regard the bottom volume and the top volume as one of the signs that the stock price will reverse.

- Time means that we need to think about the trajectory of stock price from the perspective of time, and the trend of stock price is cyclical. There are two main aspects, one is the stock price range, the other is the trend. The time here refers to the length of time that the stock price remains within a specific price range or trend and runs. The trend is difficult to change significantly in the short term, but it is not unchanged forever. As time changes, new trends will appear unconsciously.

- Space refers to the size and range of stock price changes, or the range of stock prices. Every stock has a high point and low point that can be reached due to the influence of the company's operation and market factors. When the stock reaches a high point or the low point reverses, the price between the high and low points is the size of the stock's operation space.

Common technical indicators and classification are shown in Table 7.

Table 7: Important technical indicators of stock time series

Technical index category	Name and abbreviation of technical indicators	key parameter
Trend indicator	Ma simple moving average: Ma5, ma10	5, 10
	MACD exponential smooth moving average: diff, DEA, MACD	26, 12, 9
	MTM dynamic index: MTM	6, 6
	Trend indicator: DMI	14, 6
	Index average index: expma	12, 50
	Bias deviation rate: bias	12
Reverse trend indicator	KDJ random index: K, D, j	9, 3, 3
	ROC change rate: ROC	12, 6
	RSI relatively strong index: RSI	6
Pressure support index	Boll brin belt	26, 2
Volume price index	Sobv energy tide: sobv	1
	Wvad William variation dispersion: wvad	24, 6
Volume index	Standard deviation of vstd trading volume: vstd	10
	Turnover rate: turn	1
Swing index	RC change rate index: RC	8
Fluctuation index	STD standard deviation: STD	26

3. Based on establishment of time series information table for extreme point extraction

After the above analysis, the information table of extreme points based on trend division includes the following contents:

- The decision variables are given by the algorithm proposed in this paper.

- The state variables include various basic data and technical indicators de-scribed above, and the time series information table shown in Table 8 can be obtained.

Table 8: Time series information table

time	Status variable (basic data)						State variables (technical indicators)						Decision variable
minute	Opening price	Maximum price	Lowest price	Closing price	volume	Turnover	MA	MACD	Bias	KDJ	RSI	BOL	
|	|	|	|	|	—	—	—			|	|	|	+
|	|	|	|	|	—	—	—			|	|	|	-

The decision table with different time granularity contains 12 state variables, including 6 basic data and 6 technical index data, which are opening price, maximum price, minimum price, closing price, trading volume, trading volume, Ma, MACD, bias, KDJ, RSI and BOL. The extreme points extracted are used as decision variables.

The following table shows the number of extreme points after processing at different time granularity.

Table 9: Extreme points under different granularity

data sources	Number of extreme points
Number of 5-minute interval data	16911
Number of 15-minute interval data	5707
Number of 30-minute interval data	2798
Number of 60-minute interval data	1405
Number of daily interval data	381

Based on the analysis of ARIMA model, this section proposes a multi granularity time series decision table conversion method based on trend division. By using the method of time series value recursion, the extreme points of time series are extracted, and the extreme points are taken as decision variables, and the time series decision table is constructed by combining the relevant technical indicators of stock time series and other data. Transforming time series data into decision table is the basis of time series classification modeling. Therefore, this chapter is a key content of this paper, paving the way for the next application.

4 Validation of time series decision table based on support vector machine

This experiment uses Python language sklearn. The SVC method in the SVM module is used to implement the support vector machine model. The kernel function is set to "linear" by the kernel, and the data set division method in the sklearn module is also used to realize the data set division. The data set is divided according to the ratio of 7:3.

Table 10 shows the accuracy of support vector machine model:

Table 10: Accuracy of support vector machine models with different granularity

data set	Accuracy
5-Minute interval dataset	77.82%
15-Minute interval dataset	79.07%It is necessary to
30-Minute interval dataset	76.66%
60-Minute interval dataset	76.30%It is necessary to

Table 11 shows the confusion matrix generated by support vector machine models with different granularity:

Table 11: Confusion matrix of support vector machine models with different granularity

			Predict	
			0	1
5-Minute interval support vector machine	TRUE			
		0	1980	560
		1	565	1969
			Predict	
			0	1
15-Minute interval support vector machine	TRUE			
		0	685	192
		1	167	669
			Predict	
			0	1
30-Minute interval support vector machine	TRUE			
		0	320	89
		1	107	324
			Predict	
			0	1
60-Minute interval support vector machine	TRUE			
		0	164	54
		1	46	158

Accuracy of different granularity support vector machine models with different step sizes removed:

Table 12: Different SVM models with different step removed

data set	Remove the accuracy of step size 1	Remove the accuracy of steps 1 and 2	Remove the accuracy of steps 1, 2 and 3
5Minute line dataset	88.58%	91.03%	93.64%
15Minute line dataset	88.48%	90.25%	93.12%
30Minute line dataset	89.91%	90.02%	92.31%
60Minute line dataset	86.87%	88.44%	91.45%

The confusion matrix generated by support vector machine models with different granularity is shown in the table below:

Table 13: 5-minute dataset

			Predict	
Remove the support vector machine with step size of 1	TRUE		0	1
		0	1489	191
		1	196	1515
			Predict	
Remove support vector machines with steps of 1 and 2	TRUE		0	1
		0	1160	120
		1	111	1177
			Predict	
Remove support vector machines with steps of 1, 2 and 3	TRUE		0	1
		0	908	63
		1	69	1028

Table 14: 15-minute dataset

			Predict	
Remove the support vector machine with step size of 1	TRUE		0	1
		0	498	66
		1	65	509
			Predict	
Remove the support vector machine with steps of 1 and 2	TRUE		0	1
		0	401	41
		1	44	381
			Predict	
Remove support vector machines with steps of 1, 2 and 3	TRUE		0	1
		0	341	28
		1	21	307

Table 15: 30-minute dataset

		Predict	
Remove the support vector machine with step size of 1	TRUE	0	1
	0	254	24
	1	33	254
		Predict	
Remove support vector machines with steps of 1 and 2	TRUE	0	1
	0	199	26
	1	17	186
		Predict	
Remove the support vector machines with steps of 1, 2 and 3	TRUE	0	1
	0	138	12
	1	14	172

Table 16: 60-minute dataset

		Predict	
Remove the support vector machine with step size of 1	TRUE	0	1
	0	133	20
	1	17	112
		Predict	
Remove support vector machines with steps of 1 and 2	TRUE	0	1
	0	92	11
	1	14	97
		Predict	
Remove the support vector machines with steps of 1, 2 and 3	TRUE	0	1
	0	88	8
	1	7	70

From the above data, it is known that using multi-granularity time series decision table based on trend partition and related classification methods can effectively predict the trend points of stock time series. The accuracy of prediction is high, and no fitting phenomenon will occur. Better prediction results can be obtained by adjusting tolerance errors and filtering extreme points. The specific mechanism will be explored in future research.

5   Conclusions

This paper proposed a new time series trend division method, which uses tol-erance error (or step adjustment) to determine the trend division of time series. This method can process time series into two modes, rise and fall. Based on the trend partitioning method of time series with different granularity, the series are divided into rise and fall trends, which avoid the complexity of threshold, numerical length and multiple trends. Take the easiest way to identify trends. The method is simple and the time complexity is low. On the other hand, with time granularity as the main measure, the nested structure of time series with different time granularity based on trend partitioning is obtained. This structure can fully express the struc-ture characteristics of time series, which provides a good basis for further study on time series.

A trend-based extreme point extraction algorithm is proposed, which can ex-tract extreme points with different time granularity. On the basis of getting the extreme points, the high (1) and low (0) extreme points are used as decision vari-ables. The other information (such as mean, volume, etc.) of the extreme points is extracted to establish time series decision tables that divide different

time granu-larities based on trends. On this basis, the decision table was validated by random forest method, and the results were good.

The proposed multi-granularity time series decision table can accurately characterize the data characteristics of time series and can be used for complex time series mining methods. The data of decision table is symmetrical, does not produce unbalanced data phenomenon of general decision problem, has higher mining ac-curacy, and significantly improves the prediction accuracy compared with the original time series regression and measurement model. It provides a new method for the study of time series problem, and has strong theoretical significance and application value.

References

[1] Duplakova D, Teliskova M, Duplak J, et al. Determination of Optimal Production Process Using Scheduling and Simulation Software[J]. International Journal of Simulation Modelling, 2018, 17(4): 609–622.

[2] Janekova J, Fabianova J, Fabian M. Assessment of Economic Efficiency and Risk of The Project Using Simulation[J]. International Journal of Simulation Modelling, 2019(2): 18.

[3] Junior W S, Montevechi J, Miranda R, et al. Economic Lot-size using Machine Learning, Parallelism, Metaheuristic and Simulation[J]. International Journal of Simulation Modelling, 2019(2): 18.

[4] Xu, W.; Sun, H.Y.; Awaga, A.L.; Yan, Y.; Cui, Y.J. (2022). Optimization approaches for solving production scheduling problem: A brief overview and a case study for hybrid flow shop using genetic algorithms, *Advances in Production Engineering & Management*, Vol. 17, No. 1, 45-56

[5] Lu Zhang, Yan Yan, Wei Xu, Jun Sun, Yuanyuan Zhang, "Carbon Emission Calculation and Influencing Factor Analysis Based on Industrial Big Data in the "Double Carbon" Era", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2815940, 12 pages, 2022.

[6] Agrawal R., Faloutsos C, and Swami A. Efficient similarity search in sequence databases[Z]. In D. Lomet, editor, Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms (FODO), pages69-84, *Chicago, Illinois*, 1993. Springer Verlag.

[7] Burrus C S, Gopinath R A., and Guo H. Introduction to wavelets and wavelet transforms[M] A Primer. *Prentice Hall*, 1998.

[8] Box G E P, Jenkins G M, Reinsel G C. Time series analysis: forecasting and control[M]. *Wiley.com*, 2013

[9] Granger C W J, Andersen A P. An introduction to bilinear time series models[M]. *Göttingen: Vandenhoeck und Ruprecht*, 1978

[10] Hossain E, Hossain M S, Zander P O, et al. Machine learning with Belief Rule-Based Expert Systems to predict stock price movements [J]. *Expert Syst Appl*, 2022, 206(13).

[11] Ji G, Yu J M, Hu K, et al. An adaptive feature selection schema using improved technical indicators for predicting stock price movements [J]. *Expert Syst Appl*, 2022, 200(12).

[12] Li G Z, Zhang A N, Zhang Q Z, et al. Pearson Correlation Coefficient-Based Performance Enhancement of Broad Learning System for Stock Price Prediction [J]. *IEEE Trans Circuits Syst II-Express Briefs*, 2022, 69(5): 2413-7.

[13] Deng, C. R., et al. (2022). "Multi-step-ahead stock price index forecasting using long short-term memory model with multivariate empirical mode decomposition." *Information Sciences* 607: 297-321.

[14] Gao, R. Z., et al. (2022). "Forecasting the overnight return direction of stock market index combining global market indices: A multiple-branch deep learning approach." *Expert Systems with Applications* 194: 18.

[15] Gao, Z., et al. "Financial sequence prediction based on swarm intelligence algorithms and internet of things." *Journal of Supercomputing*: 21.

[16] Gupta, U., et al. (2022). "StockNet-GRU based stock index prediction." Expert Systems with Applications 207: 16.

[17] He Q Q, Siu S W I, Si Y W. Instance-based deep transfer learning with attention for stock movement prediction [J]. *Appl Intell*, 22.

[18] Kanwal A, Lau M F, Ng S P H, et al. BiCuDNNLSTM-1dCNN-A hybrid deep learning-based predictive model for stock price prediction [J]. *Expert Syst Appl*, 2022, 202(15).

[19] Kumar R, Kumar P, Kumar Y. Three stage fusion for effective time series forecasting using Bi-LSTM-ARIMA and improved DE-ABC algorithm [J]. *Neural Comput Appl*, 17.

[20] Li R R, Han T, Song X. Stock price index forecasting using a multiscale modelling strategy based on frequency components analysis and intelligent optimization [J]. *Appl Soft Comput*, 2022, 124(15).

[21] Liang M X, Wu S C, Wang X L, et al. A stock time series forecasting approach incorporating candlestick patterns and sequence similarity [J]. *Expert Syst Appl*, 2022, 205(26).

[22] Wang C J, Chen Y Y, Zhang S Q, et al. Stock market index prediction using deep Transformer model [J]. *Expert Syst Appl*, 2022, 208(10).

[23] Preis, T., Schneider, J. J. & Stanley, H. E. Switching processes in financial markets. *Proceedings of the National Academy of Sciences of the United States of America* 108, 7674–8 (2011).

C O P E

Member since 2012
JM08090

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:
Yang, H.; Gao, X.; Han, L.; Cui, W. (2022). A financial time series data mining method with different time granularity based on trend Division, *International Journal of Computers Communications & Control*, 17(6), 4952, 2022.
https://doi.org/10.15837/ijccc.2022.6.4952