ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20230224000379
Original scientific paper

A Research on Dimension Reduction Method of Time Series Based on Trend Division

Haining YANG*, Xuedong GAO, Wei CUI

Abstract: The characteristics of high dimension, complexity and multi granularity of financial time series make it difficult to deal with effectively. In order to solve the problem that the commonly used dimensionality reduction methods cannot reduce the dimensionality of time series with different granularity at the same time, in this paper, a method for dimensionality reduction of time series based on trend division is proposed. This method extracts the extreme value points of time series, identifies the important points in time series quickly and accurately, and compresses them. Experimental results show that, compared with the discrete Fourier transform and wavelet transform, the proposed method can effectively process data of different granularity and different trends on the basis of fully preserving the original information of time series. Moreover, the time complexity is low, the operation is easy, and the proposed method can provide decision support for high-frequency stock trading at the actual level.

Keywords: data dimensionality reduction; Fourier transform; time series; trend division; wavelet transform

1 INTRODUCTION

Before the actual data mining operation, we need to solve a basic problem: data representation and preprocessing. Because time series is usually a very long series, the representation of data is particularly important for time series, and it is extremely difficult to directly operate a continuous, high-dimensional data space [1-4]. For example, when hyperspectral data is measured in the field of remote sensing, problems such as high data redundancy and dimensional disaster will occur, resulting in large amount of calculation and poor classification effect in remote sensing data processing [5-7].

Therefore, it is necessary to reduce the dimension of the time series, to represent it in a continuous or discrete form, and then to study the new data [8]. Data dimension reduction is not only the main means to study the characteristics of time series, but also the most common way to characterize time series data. Generally speaking, how to achieve the balance between reducing and preserving the important trend features contained in the original sequence is very important for the dimensional reduction of time series [9].

Financial time series data present the characteristics of massive and high-dimensional, and its highly nonlinear, non-stationary, volatile and noisy data bring great troubles to the prediction of financial time series. Meanwhile, multi-time granularity and multi-data mode make the study of financial time series data more complex and diversified. Currently, commonly used dimensionality reduction techniques include principal component analysis and singular value decomposition, discrete Fourier transform, wavelet transform, PAA, sax, etc. [10]. However, these are all for a single time scale [11], and time series data of different granularity cannot be processed at the same time, resulting in overfitting. Therefore, in order to master the change trend and important features of time series, it is a very effective dimensionality reduction method to represent the series by important points in the series [12].

In order to solve the problem that the commonly used dimensionality reduction methods cannot reduce the dimensionality of time series with different granularity at the same time, in this paper, a method for dimensionality reduction of time series based on trend division is proposed. This method extracts the extreme value points of time series, identifies the important points in time series quickly and accurately, and compresses them. Experimental results show that, compared with the discrete Fourier transform and wavelet transform, the proposed method can effectively process data of different granularity and different trends on the basis of fully preserving the original information of time series. Moreover, the time complexity is low, the operation is easy, and the proposed method can provide decision support for high-frequency stock trading at the actual level.

2 RELATED WORK

2.1 Dimension Reduction

The main purpose of dimension reduction is to reduce the dimension of space, project an n-dimensional time series into a k-dimensional space, so that the process distance should be kept as much as possible, and use an index technology to realize retrieval in a new space [13]. When searching, the query s is mapped to a new space, a time series is represented as a k-dimensional point, and the nearest neighbor to s is found in the new space $k \ll n$.

Let f denote the dimensionality reduction technique and the best f transformation requirement. Otherwise, the following two situations will occur: $D(F(X), F(Y)) = D(X, Y)$.

When false dismissals occur, that is, some originally similar sequences are not retrieved at all; $D(F(X), F(Y)) > D(X, Y)$.

When false alarms occur, that is, the retrieved similar sequences are not necessarily similar. $D(F(X), F(Y)) < D(X, Y)$.

Therefore, in order to ensure the correctness of the final results, we must ensure that there is no omission. If yes, there must be, which ensures that the sequence retrieved in the new space contains all the correct results, and there will be no false positives, but there may be false positives, which can be eliminated by further post-processing [14]. The commonly used dimension reduction techniques are: discrete Fourier transform, wavelet transform, PAA, sax, etc. $D(X, Y) < \varepsilon$ $D(F(X), F(Y)) < \varepsilon$.

2.2 Discrete Fourier Transform

For a sequence of n points, its discrete Fourier transform (DFT) is: $X = x_1, x_2, \ldots, x_n$,

$$X_f = \frac{1}{\sqrt{n}} \sum_{t=1,2,\ldots,n} x_f \mathrm{e}^{-j2\pi ftln} \quad \left(f = 1,2,\ldots,n;\ j^2 = -1\right).$$

The time complexity of DFT is [15]. Fourier transform is a pure frequency domain analysis method, which reflects the overall frequency characteristics of the whole signal at all times, and cannot provide the frequency characteristics at local time [16]. Fourier transform transforms the signal from time domain to frequency domain, but it cannot combine the two organically. This is because the Fourier transform is an integral in the whole time domain, and the time domain waveform of the signal does not contain any frequency domain information, so it has no function of localizing and analyzing the signal. In other words, it is impossible to know when a certain frequency is generated. Windowed Fourier transform multiplies a time window function and the function to be analyzed, and then carries out Fourier transform. The result can describe the information in a local time period, but for a time-varying unstable signal, it is difficult to find a suitable time window for different time periods. $O(n\log n)$.

2.3 Discrete Wavelet Transform

Wavelet analysis method is developed on the basis of Fourier method [17]. It reflects the difference of signal in time domain and frequency domain at the same time. It has good localization properties in both time domain and frequency domain, and can decompose all kinds of mixed signals with different frequency composition intertwined into block signals in different frequency bands [18]. It has the characteristics of multi-scale and time shift invariance. Wavelet transform is a non-stationary signal analysis method. It represents or approximates a function through the translation and expansion of a basic wavelet function that satisfies the conditions. $\int_R \psi(x)\mathrm{d}x = 0 \quad \psi(x)$.

Discrete wavelet transform (DWT) divides the time series into scale part and detail part. The scale part is obtained by convolution low-pass filter of the sequence to be analyzed, which reflects the general trend and direction of the original sequence. The detail part is obtained by convolution high pass filter of the signal to be analyzed, indicating the difference in detail of the signal [19]. Further implement DWT on the scale part to get more detailed scale part and detail part. This process can continue all the time, so wavelet transform has the characteristics of multi-scale decomposition. DWT is applied to a sequence with length n to obtain two scale sequences and detail sequences with length $n/2$. In this way, the length of the scale sequence is reduced to 1/2 of the original signal length every time DWT is performed. If we carry out the wavelet transform with scale 3, the length of the scale sequence will be 1/8 of the length of the original sequence. The higher the number of scales, the shorter the length of the scale sequence, and the fuzzier the signal. The scale sequence tracks the trend of the original sequence well, so the scale sequence can be used to represent the original sequence, achieving a substantial reduction in the amount of data, while the amount of information is lost relatively less.

3 METHODOLOGY

3.1 Method Introduction

After studying various dimensionality reduction methods and observing the characteristics of financial time series, this paper proposes a new dimensionality reduction method: trend division. The main idea is: a trend must have extreme points [20], a local maximum extreme point and two local minimum extreme points before and after constituting 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.

In this paper, we use Python to write programs to find the extreme points of time series, and then identify the rising and falling trends in time series. In order to verify the feasibility of the above algorithm, we randomly select a group of time series data to extract the extreme points.

Taking the time series of [0, 6, 25, 23, 2, 20, 15, −8, 15, 3, 1, 8, 0, 4, −15, −3, 5, 4, 8, 2, 13, 8, 10, 3, −1,20, 7, 3, 0] 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 the following Fig. 1:

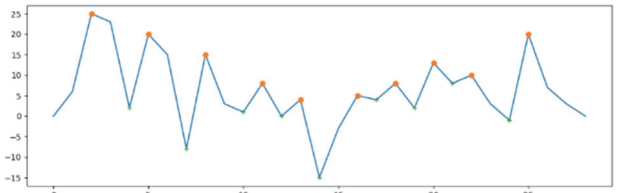

Figure 1 Schematic diagram of extreme point identification

3.2 Trend Division under Different Granularity

This paper selects the closing price data of the 5-minute, 10-minute, 15-minute, 60-minute and daily of 399006.sz gem index from August 21, 2019 to August 19, 2022. Taking the above data as experimental data, extreme points were extracted, and the results are as follows.

Table 1 Basic data

Data description	Number of data	Number of extreme points
5 minutes	34944	16911
15 minutes	11648	5707
30 minutes	5816	2798
60 minutes	2912	1405
Daily data	728	381

Support vector machine is used to predict the data set with mixed time granularity, and the results are as follows:

Table 2 Basic data

Data set	Accuracy
Mixed time granularity (including daily, 60-minute, 30-minute, 15-minute and 5 -minute)	63.88%
Remove steps 1, 2, 3	83.50%

Table 3 Confusion model

Predict				
Support vector machines with mixed time granularity	True		0	1
		0	2144	1446
		1	1130	2358
Predict				
Remove the support vector machines with steps 1, 2 and 3	True		0	1
		0	3827	536
		1	729	2576

4 EXPERIMENTAL COMPARISONS OF THREE DIMENSIONALITY REDUCTION METHODS

This paper selects 34944 pieces of closing price data of the 15 minutes of 399006.sz gem index from August 21,

2019 to August 19, 2022 as experimental data. Three dimensionality reduction methods are used to reduce the dimension of data and check the effect.

4.1 Discrete Fourier Transform

The Fourier transform takes a function and creates a series of sine waves. Combining these sine waves will keep approaching the original function [8]. It can be seen from the above results that the more components of the discrete Fourier transform are used, the closer the approximation function is to the actual stock price (100 component transforms are almost the same as the original sequence). Therefore, we know that Fourier transform can be used to extract long-term and short-term trends.

Figure 2 Fitting diagram with Fourier component of 3

Figure 3 Fitting diagram with Fourier component of 6

Figure 4 Fitting diagram with Fourier component of 9

Figure 5 Fitting diagram with Fourier component of 50

Figure 6 Fitting diagram with Fourier component of 100

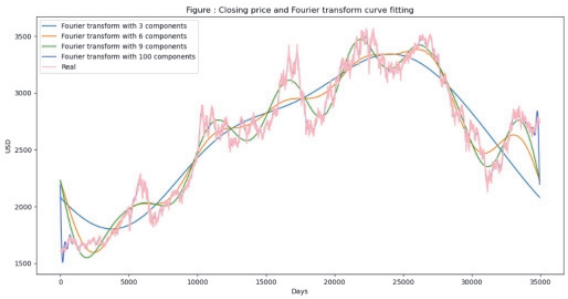

Figure 7 Fitting diagram of different Fourier components

Table 4 Fitting evaluation index results when the Fourier component is 100

Evaluating indicator	MSE	RMSE	Mae:	MAPE
Numerical value	1397.7261	37.3861	24.4260	0.0095

When the component transformation with small proportion such as 3, 6 and 9 is adopted, the long-term transformation trend of time series can be well seen, but the judgment of local trend is also lost. When 50100 and other large component transformations are used, the fitting of the data is better by naked eye observation, and the fitting effect is judged by the evaluation index, with large error. According to the analysis, because Fourier transform changes the representation of time series, the final fitting does not use the original data, that is, it ignores the important points in the original data.

4.2 Discrete Wavelet Transform

Daubechies2 and daubechies8 wavelets were selected for filtering and dimensionality reduction during the experiment. The results are as follows:

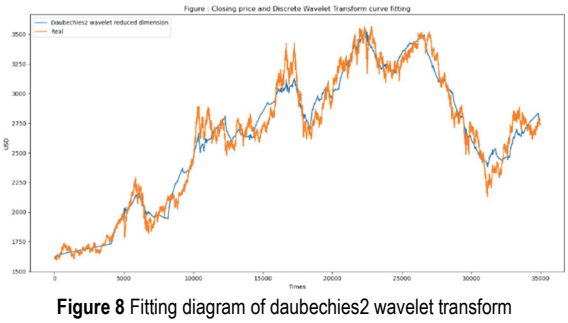

Figure 8 Fitting diagram of daubechies2 wavelet transform

Compared with Fourier transform, wavelet transform has better fitting effect on local trend, and daubechies8 wavelet is obviously better than daubechies2 wavelet. Haar wavelet is also used in the experiment, but the algorithm of Haar wavelet cannot recognize some data in the data set, and there is a calculation interruption in the middle, so this method also has disadvantages.

Figure 9 Fitting diagram of daubechies8 wavelet transform

Table 5 Fitting evaluation index results of daubechies8

Evaluating indicator	MSE	RMSE	Mae:	MAPE
Numerical value	4476.2030	66.9044	53.8978	0.0206

4.3 Micro Trend Identification

(1) Extract extreme points

In order to measure the dimensionality reduction of the original data, the data compression ratio is selected as the evaluation index, and the calculation method is as follows:

$$CR(\%)=\frac{w}{m}\times 100$$

where, it represents the amount of data in the original time series and the number of key points after compression. Obviously, the smaller the compression rate, the greater the degree of dimensionality reduction for the data.

The local extreme points of 34944 data were identified, and the step of extreme point identification was set to 1. After reduction, 16911 extreme points, 8455 maximum points and 8456 minimum points can be obtained, and the compression ratio is 48%. The results are as follows:

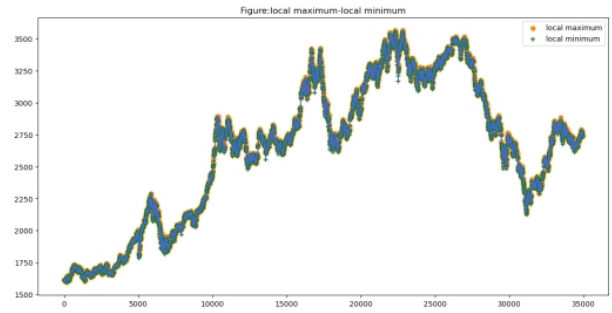

Figure 10 Extraction of extreme points of 15-minute

Because the 15-minute *s* are too dense, the extreme points of 728 daily s from August 21, 2019 to August 19, 2022 are extracted to observe the effect.

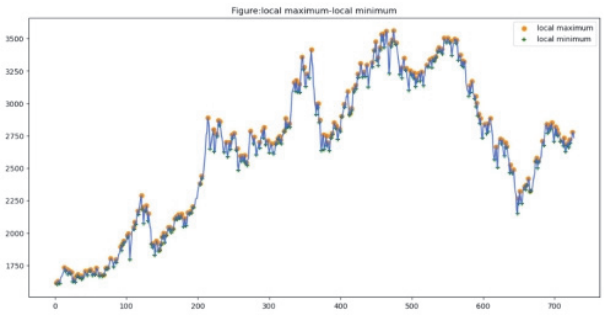

Figure 11 Extreme point extraction of daily

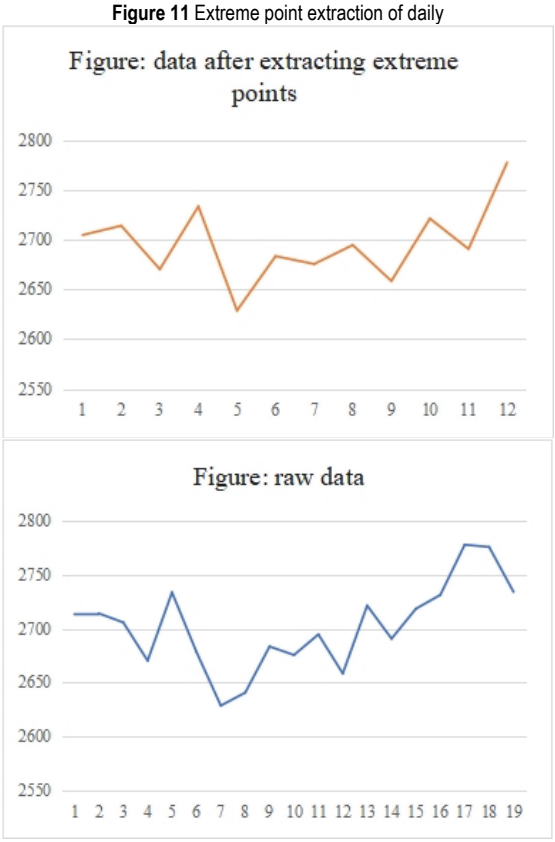

Figure 12 Comparison of two method

The last 20 pieces of data are enlarged and visualized. The left figure shows the original data. It can be seen that there are many fluctuation points in the original time series, and these fluctuation points can be combined to reduce the amount of data required. The right figure shows the dimension reduction data after micro trend recognition. It can be seen that the new sequence can well represent the important extreme points of the original sequence, and greatly reduce the amount of data required, which is conducive to improving the calculation speed and accuracy. In addition, it is worth emphasizing that in this step, the reduction is based on finer granularity, and the purpose is to identify those key points. If the granularity is scaled to a coarser level and then reduced, those important extreme points may be lost. As shown in the above figure, extreme points usually do not last long. If you scale to a coarser granularity, you need to average some data near the extreme points, which may cause important extreme point information to be submerged.

(2) Verify the prediction effect

In the following experiments, we establish time series decision tables based on different indicators, and use support vector machines to verify the effect of dimensionality reduction. Then, the extreme points with step size of 1, 1 and 2, and 1, 2 and 3 are removed, and the data are fitted and predicted. The results are as follows.

Table 6 Dimension reduction compression results

Data set	Number of extreme points	Compression ratio
Remove the sequence with step size of 1	8451	24.18%
Remove the sequence with steps of 1 and 2	5637	16.13%
Remove the sequence with steps of 1, 2 and 3	4228	12.09%

Table 7 Accuracy

Data set	15-minute dataset	Remove the accuracy of step size 1	Remove the accuracy of steps 1 and 2	Remove the accuracy of steps 1, 2 and 3
Accuracy	79.07%	76.48%	89.25%	92.12%

When the step size is removed as 1, the reason for the reduction of accuracy is that some effective points may be misjudged. Therefore, the method of identifying extreme points needs to be improved in the future.

4.3 Time Complexity Comparison

The time complexity of the algorithm is a function, which qualitatively describes the running time of the algorithm. From the perspective of time complexity, the lower the time complexity is, the faster the calculation speed is.

The time complexity of DFT is. The difference between wavelet transform and Fourier transform is that the transformed time series are still in the same time domain, with time domain localization, and the time complexity of the coefficient generation algorithm is O (n), which is more effective than DFT. Therefore, the computing speed is slow, and it is unable to quickly process a large number of financial time series with different granularity. $O(n\log n)$.

The time complexity of the time series dimensionality reduction method based on trend division is 1. It only needs to run the time series from beginning to end without any transformation. It directly compares the values to get the local trend, which greatly improves the running speed, and can process the time series with different granularity at the same time. Therefore, from the perspective of algorithm, the method proposed in this paper is faster.

5 SUMMARY

In this paper, the traditional dimensionality reduction methods (Discrete Fourier transform and wavelet transform) and the proposed dimensionality reduction method based on extreme point recognition are compared experimentally. It is considered that the method proposed in this paper is better in identifying local trends and extreme points of time series. The specific conclusions are as follows:

(1) It can extract extreme points with different granularity and different trends.

Fourier transform and wavelet transform change the representation of the original data and compress and filter the data, so they can only process the data with a unified time width. The dimensionality reduction method proposed in this paper is to extract the important extreme points in the original data that are effective for subsequent work, and then compress the time series. Therefore, it can process data with different granularity and different trends at the same time, which greatly shortens the time of dimension reduction.

(2) Able to handle large amounts of data.

Traditional dimensionality reduction methods have high requirements for computers and cannot process a large number of data. Fourier transform is difficult to deal with sequences of different lengths and does not support weighted distance measurement. Wavelet transform also does not support weighted distance measurement. The method proposed in this paper can process a large number of data with different granularity and different trends in a short time.

(3) More targeted data processing.

Fourier transform and wavelet transform can only transform the data to form and extract the trend of data transformation, but cannot make accurate judgment for the data at a certain time point. The dimension reduction method based on extreme point identification proposed in this paper is to reduce the dimension of local minima, which retains the original data useful for subsequent work.

(4) The time complexity is low and the model is simple.

The time complexity of Fourier transform and wavelet transform are $o(n)$ and $O(n)$, respectively. It is unable to rapidly process a large number of financial time series with different granularity at the same time. From the perspective of the model, the complexity is high. The time complexity of the dimensionality reduction method proposed in this paper is 1, which can greatly improve the calculation speed, and the model is simple and convenient for practical operation. $O(n\log n)$.

6 REFERENCES

[1] Cai, Y. X. & Gong, G. Y. (2021). The Gold Price and the Economic Policy Uncertainty Dynamics Relationship: The Continuous Wavelet Analysis. *Economic Computation and Economic Cybernetics Studies and Research*, *55*(1), 105-116. https://doi.org/10.24818/18423264/55.1.21.07

[2] Zhang, L., Li, Y., Wu, X., & Feng, Q. (2021), Research on Fractal Portfolio Model under Power-Law Distribution of Return Rate. *Economic Computation and Economic Cybernetics Studies and Research*, *55*(1), 219-232. https://doi.org/10.24818/18423264/55.1.21.14

[3] Zhang, L., Ma, J., Liu, X., Zhang, M., Duan, X., & Wang, Z. (2022). A Novel Support Vector Machine Model of Traffic State Identification of Urban Expressway Integrating Parallel Genetic and C-Means Clustering Algorithm. *Tehnicki vjesnik - Technical Gazette*, *29*(3), 731-741. https://doi.org/10.17559/TV-20211201014622

[4] Fazal, M. Z., Khan, S., Abbas, M. A., Nawab, Y., & Younis, S. (2021). Machine Learning Approach for Prediction of Crimp in Cotton Woven Fabrics. *Tehnicki vjesnik - Technical Gazette*, *28*(1), 88-95. https://doi.org/10.17559/TV-20191018180716

[5] Vanaga, R. & Sloka, B. (2020). Financial and capital market commission financing: aspects and challenges. *Journal of Logistics, Informatics and Service Science*, *7*(1), 17-30. https://doi.org/10.33168/LISS.2020.0102

[6] Shanmugathas, S. & Ashoka, K. (2019). Material sourcing in a strategic way: Evaluation of consequences on the organizational performance. *Journal of Logistics, Informatics and Service Science*, *6*(1), 69-86.

[7] Kim, S., Park, S., & Chu, W. (2001). An index-based approach for similarity search supporting time warping in large sequence databases. *Proceedings 17th International Conference on Data Engineering*, 607-614. https://doi.org/10.1109/ICDE.2001.914875.

[8] Agrawal, R., Lin, K. I., Sawhney, H. S., & Shim, K. (1995). Fast similarity search in the presence of noise, scaling, and translation in times series databases. *VLDB*, 490-501.

[9] Shatkay, H. (1995). *The Fourier transform - a primer*. Technical report cs-95-37, Department of computer science, Brown University.

[10] Manmatha, R. & Rath, T. M. (2003). Indexing of handwritten historical documents - recent progress. *Proceedings of the 2003 Symposium on document image understanding technology (SDIUT)*, 77-85.
[11] Arneodo, A., Manneville, S., & Muzy, J. F. (1998). Toward log normal statistics in high Reynolds number turbulence. *The European Physical Journal B*, *1*(1), 129-140.
[12] Agrawal, R., Lin, K., Sawhney, H. S., & Shim, K. (1995). Fast similarity search in the presence of noise, scaling, and translation in time series databases. *Proceedings of the 21st VLDB conference Zurich*, 490-501.
[13] Berthold, M. R. & Hppner, F. (2016). On clustering time series using Euclidean distance and person correlation. *Expert systems with applications*, *52*(6), 26-38.
[14] Rath, T. M. & Manmatha, R. (2002). *Lower bounding of dynamic time warping distances for multivariate time series Technical report MM-40*. Center for intelligent information retrieval, University of Massachusetts Amherst.
[15] Stollnitz, E., DeRose, T., & Salesin, D. (1995). Wavelets for computer graphics a primer. *IEEE Computer graphics and applications.*
[16] Stadnik, B., Raudeliūnienė, J., & Davidavičienė, V. (2016). Fourier analysis for stock price forecasting: assumption and evidence. *Journal of Business Economics and management*, *17*(3), 365-380
[17] Cunha, C. F. F. C., Carvalho, A. T., Petraglia, M. R., & Lima, A.C.S. (2015). A new wavelet selection method for partial discharge denoising. *Electric Power Systems Research*, *185*, 184-195. https://doi.org/10.1016/j.epsr.2015.04.005
[18] Azami, H., Mohammadi, K., Mohammadi, K., & Bozorgtabar, B. (2012). An improved signal segmentation using moving average and Savitzky-Golay filter. *Journal of signal and information processing*, *3*(1), 39-44. https://doi.org/10.4236/jsip.2012.31006
[19] Lu, W., Rui, H., Liang, C., Jiang, L., Zhao, S., & Li, K. (2020). A Method Based on GA-CNN-LSTM for Daily Tourist Flow Prediction at Scenic Spots. *Entropy*, *22*(3), 261. https://doi.org/10.3390/e22030261
[20] Passricha, V. & Aggarwal, R. (2020). A comparative analysis of pooling strategies for revolutionary neural network based Hindi ASR. *Journal of ambient intelligence and humanized computing*, *11*(2), 675-691. https://doi.org/10.1007/s12652-019-01325-y

Contact information:

Haining YANG, Dr.
(Corresponding author)
Department of Management Science and Engineering,
School of Economics and Management,
University of Science and Technology Beijing,
Beijing 100083, China
E-mail: yanghaining@apiins.com

Xuedong GAO, Prof., Dr.
Department of Management Science and Engineering,
School of Economics and Management,
University of Science and Technology Beijing,
Beijing 100083, China
E-mail: gaoxuedong@manage.ustb.edu.cn

Wei CUI, Dr.
School of Economics and Management,
China University of Geosciences (Beijing),
Beijing 100083, China
E-mail: cuiw@cugb.edu.cn