A Clustering Method of Non-stationary Time Series and its Application in CSI 300 Analysis

Dongyang Ye, Kaiji Liao, Haitao Song[1*]

School of Business Administration, South China University of Technology, Guangzhou, Guangdong, 510641, China
E-mail: htsong@scut.edu.cn

***Abstract* — Correlation analysis is the basic work for discovering the inner connection between sequences and has important implications to subsequent work like classification and dynamic analysis. This paper has studied the correlation measurement method of clustering non-stationary time series. Unlike most of the studies using Pearson's correlation coefficient, we propose to calculate the correlation between stocks combing the log-return and Spearman correlation coefficient. Using this method, the stocks in CSI 300, the most popular stock indicator in China, have been clustered and we use three evaluation methods to observe the clustering effect. The result shows this cluster method can come up with a satisfactory classification. Further analysis indicates that the implicit links between stocks, such as equity relationship, are implied in clusters. In general, the correlation between stocks is measured accurately and the proposed method can solve the correlation measurement problem of clustering non-stationary time series more reasonably.**

Keywords - Correlation analysis, Time series, Spearman correlation coefficient, Cluster analysis, CSI 300

I. INTRODUCTION

Time series is an important kind of complex data, which is widespread in all areas of society. How to analyze these time series effectively and reveal useful information hidden in it is an important topic to help people understand things and make scientific decisions. Cluster analysis is an important method of unsupervised learning, whose goal is to put a limited unmarked data set into a limited set of "natural" separate data structures[1]. The clustering of time series is putting them into several categories in accordance with their similarity.

However, long-term trends of time series will affect the similarity of two series. For example, the United States market indices are related to the population of Pakistan, merely because of the upward trend in both of them[2]. Obviously, they are non-stationary time series. Stationary data represent any class of data whose statistical properties do not change over time[3]. In addition, commonly used methods for calculating correlation coefficients are only limited to the analysis of stationary time series[4]. Only eliminating the long-term trends can the real correlations between sequences be reflected. Although Fielitz[3] shows that the logarithmic return of the stock is not strictly stationary, it is commonly recognized that the price series is non-stationary and the logarithmic return series is stationary in finance research[5]. The non-strict stationary of log-return makes that the correlation between different stocks will change over time, and that will bring difficulties to the clustering of time series.

On the choice of the correlation coefficient, researchers commonly use Pearson correlation coefficient and Spearman correlation coefficient. The Pearson correlation coefficient is applied to relationship of normal distribution and interval measured variables. The Spearman correlation coefficient is the non-parameter form of Pearson correlation coefficient without the demand of normal distribution. Recently, many studies in other subjects[6,7,8,9,10] have used the Spearman correlation coefficient for the measurement of correlation between variables. However, almost all researchers in finance study use Pearson correlation coefficient which has distribution restrictions, when calculating correlation coefficients between stocks. The log-normal assumption is not consistent with all the characters of historical log-return of stocks, especially when many stock returns show positive excess kurtosis[5]. Obviously, the Pearson correlation coefficient is not an ideal choice.

This paper uses log-return and Spearman correlation coefficient to calculate the correlation between stocks. Even though the log-return series show non-stationary over time, only if their relative position (rank) is unchanged, using this method can still discover the fluctuation correlation between stocks. This paper will further comparatively analyze the benefits of using Spearman correlation coefficient, and use this method for clustering CSI 300 stocks. Most of researches did not evaluate the clustering quality. A few researchers like Jukka-Pekka Onnela et al.[11] used subgraph intensity and subgraph coherence to measure the effectiveness of clustering after graph clustering. Xiaohang Zhang et al.[12] judged the cluster validity by the following guidelines: if in a certain cluster the majority of stocks belonged to the same industry, and that the small number of "otherness" were the result of misclassification, then they used error rates to determine the cluster quality. In this paper, after clustering CSI300

stocks, we make many attempts and put forward 3 convenient evaluation methods for clustering quality.

II. THE CHOICE OF CORRELATION COEFFICIENT CALCULATION METHOD

Because the stock price series have a strong long-term trend, it is not suitable for analyzing directly. Simple return and log-return are two main methods which are ordinarily used for the measurement of stocks' similarity. They are defined as follow:

simple return:

$$r'_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{1}$$

log-return:

$$r_t = \ln \frac{P_t}{P_{t-1}} = \ln(1 + r'_t) \tag{2}$$

On the other hand, the Pearson correlation coefficient and Spearman correlation coefficient is defined as follows:

Pearson correlation coefficient: for the daily return of stock 1 and stock 2,

$$\rho_{12} = \frac{\sum (r_{1k} - \overline{r_1})(r_{2k} - \overline{r_2})}{\sqrt{\sum (r_{1k} - \overline{r_1})^2 (r_{2k} - \overline{r_2})^2}} \tag{3}$$

where r_{1k} is the return of the stock 1 at the day k, r_{2k} is the return of the stock 2 at the day k. $\overline{r_1}$ and $\overline{r_2}$ are the average value of r_{1k} and r_{2k}, respectively.

Spearman correlation coefficient: for the daily return of stock 1 and stock 2,

$$\theta_{12} = \frac{\sum (R_{1k} - \overline{R_1})(R_{2k} - \overline{R_2})}{\sqrt{\sum (R_{1k} - \overline{R_1})^2 (R_{2k} - \overline{R_2})^2}} \tag{4}$$

where R_{1k} is the rank of log-return of the stock 1 at the day k, R_{2k} is the rank of log-return of the stock 2 at the day k. $\overline{R_1}$ and $\overline{R_2}$ are the average value of R_{1k} and R_{2k}, respectively.

Obviously, when simple return r'_t tends to 0, r'_t and r_t are equivalent. The simple return series and log-return series are two very similar series. In order to observe their difference, we construct a special and amplified simple return. The constructed series is showed as Table I.

TABLE I. THE CONSTRUCTED SERIES SAMPLE

Simple return r'_t	Log-return r_t	Simple return r'_t	Log-return r_t
1	0.693	26	3.296
2	1.099	27	3.332
3	1.386	28	3.367
4	1.609	29	3.401
5	1.792	30	3.434
6	1.946	31	3.466
7	2.079	32	3.497
8	2.197	33	3.526
9	2.303	34	3.555
10	2.398	35	3.584
11	2.485	36	3.611
12	2.565	37	3.638
13	2.639	38	3.664
14	2.708	39	3.689
15	2.773	40	3.714
16	2.833	41	3.738
17	2.890	42	3.761
18	2.944	43	3.784
19	2.996	44	3.807
20	3.045	45	3.829
21	3.091	46	3.850
22	3.135	47	3.871

23	3.178	48	3.892
24	3.219	49	3.912
25	3.258	50	3.932

Using this series, we conduct two transformations as follow, and respectively compare simple return with log-return, Pearson correlation coefficient with Spearman correlation coefficient.

A. Linear transformation

We enlarge simple return r_t' and log-return r_t 100 times, and calculate the correlation coefficients of two series before and after transformations. The result is showed as Table II.

According to the result, after linear transformation, all the correlation coefficients equal 1.

B. Exponential transformation

We transform simple return r_t' and log-return r_t to $e^{r_t'}$ and e^{r_t}, and calculate the correlation coefficients of two series before and after transformations. The result is showed as Table II, too.

According to the result, for Spearman correlation coefficient, the correlation coefficients of two series before and after transformations are still 1. But for Pearson correlation coefficient, they are no longer equal to 1, and for simple return and log-return series, their correlation coefficients of two series before and after transformations have large different. Above all, for simple return and log-return, after some transformation such as exponential transformation, the correlation coefficients of two series before and after transformations are different. We can also foresee that if simple return series become larger, the difference between simple return series and log-return series will be larger. It indicates that although simple return and log-return are nearly the same when their value is very little, they are still different. Because log-return series has statistical characteristics which are easier to deal with, so we choose log-return for our study.

TABLE II. THE SUMMARY OF CORRELATION COEFFICIENTS RESULTS

	$100r_t'$	$100r_t$	$e^{r_t'}$	e^{r_t}
r_t'	1 / 1		0.353 / 1	
r_t		1 / 1		0.931 / 1

Note：the left side values of '/' are Pearson correlation coefficients.
the right side values of '/' are Spearman correlation coefficients.

For Pearson correlation coefficient and Spearman correlation coefficient, the calculation of Pearson correlation coefficient is depended on the element value of the series, while Spearman correlation coefficient is calculated by using the rank of series. Therefore, in the transformation without changing the rank of series, the Pearson correlation coefficient will change, but the Spearman correlation coefficient won't. Spearman correlation coefficient not only overlooks the normality requirement of series, but also uses the rank of series so that it can filter out some kind of external influence. Finally, we choose Spearman correlation coefficient as the measurement of correlation between stocks.

III. EMPIRICAL RESEARCH

This paper has chosen stocks in CSI 300 to study, in addition to 600958.SH (Oriental securities) whose time series is too short. All the data is from Wind financial terminal, and the range of time is from January 1st, 2009 to July 24th, 2015. Due to the large time span selected, the stocks had ex-dividend and their price changed, but the actual cost does not change. So the stock price and trading volume need to be fixed. The selected stock price in this research is daily closing price after forward rehabilitation.

Forward rehabilitation means keeping the existing price unchanged, cut down the previous price to make the K-line before rehabilitation pan down so that it can maintain the continuity of stock. According to the equation (2), the daily logarithmic return of stock is calculated by daily closing price and the daily closing price on previous trading day after forward rehabilitation. Then the Spearman correlation coefficients between stocks are calculated and Fig. 1 shows the frequency distribution of the correlation coefficients. In Fig. 1, the self-correlation coefficients have been removed. Fig. 1 indicates that the Spearman correlation coefficients between stocks are concentrated in the vicinity of 0.36, which implies that most of the relationship between stocks are weak. There are only 333 correlation coefficients (0.0075%) greater than or equal 0.7.

To the clustering in next step, it needs to change correlation coefficients into distance between stocks. An appropriate function is [13]:

$$d_{ij} = \sqrt{2(1 - \theta_{ij})} \quad (5)$$

After this transformation, the distance **Error! Reference source not found.** between stocks will meet the three conditions of Euclidean distance: ① $d_{ij} = 0$ if and only if i=j; ② $d_{ij} = d_{ji}$; ③ $d_{ij} \le d_{ik} + d_{kj}$. Now, the distance matrix can be used for the next step.

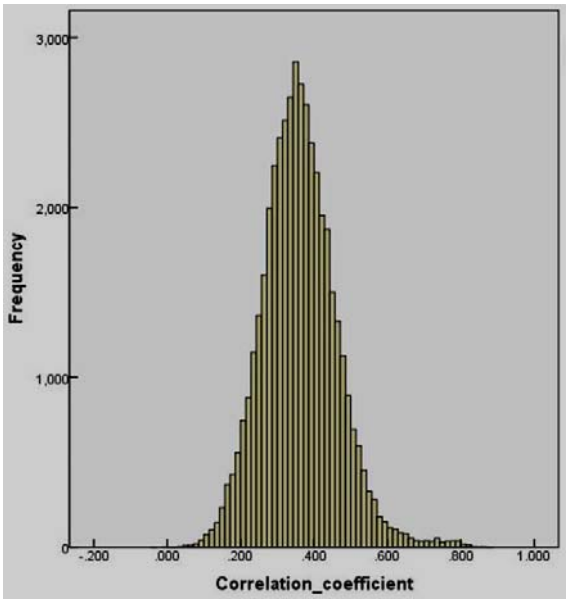

Fig.1 The Frequency Distribution of Spearman Correlation Coefficients Between Stocks

Then the hierarchical cluster analysis is conducted, based on the distance matrix. The clustering method is Ward's minimum variance method. After clustering, the dendrogram of hierarchical clustering results can be acquired, as shown in Fig. 2.

In order to make stocks separate as far as possible, without classification that is too small, the stocks are divided into 21 clusters according to Fig. 2. After that, Table III shows the brief analysis of cluster result. Then we propose 3 convenient evaluation methods for clustering quality.

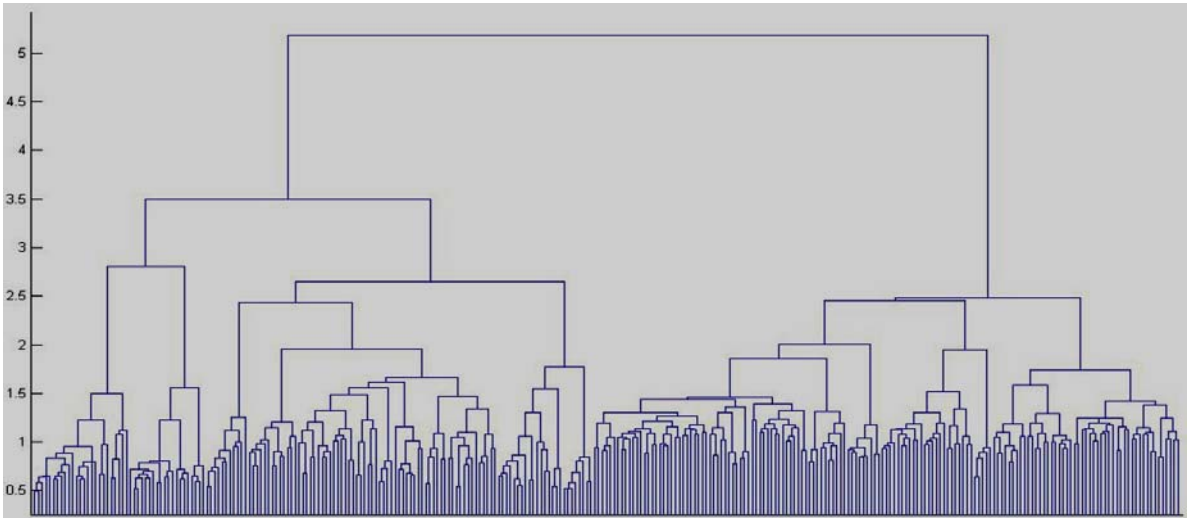

Fig.2 The Dendrogram of Hierarchical Clustering Results

TABLE III. SUMMARY OF CLUSTERING RESULT

Cluster	Main industry of the cluster	Quantity	Quantity of 'Outliers'
1	Medical Industry	17	0
2	Food Industry	7	1
3	Mineral Industry(Gold)	3	0
4	Mineral Industry(Non-ferrous metals, Non-gold)	14	0
5	Aviation Industry	4	0
6	Large State-owned Enterprises	21	0
7	Insurance Industry	4	0
8	Banking Industry	16	0
9	Media Industry	9	0
10	Concept of Internet+ Stocks	12	0
11	Steel Industry	8	0
12	States-owned Heavy Industry	19	0
13	Automation and information processing industry	28	8
14	Coal Industry	8	0
15	Appliances and Automotive Industry	10	0
16	Basic Industry	56	10
17	Liquor Industry	5	0
18	Electric Power Industry	13	0
19	Aviation Manufacturing and Shipbuilding Industry	9	0
20	Real Estate Industry	11	0
21	Securities Industry	25	3

C. sualization of clustering results

Obviously, the space which is constituted by the distance relationship between stocks is high-dimensional. In order to observe the result of classification visually, a multidimensional scaling analysis is carried out. Multidimensional scaling analysis is a multivariate statistical analysis method which is used for studying the similarity or difference between samples. It simplifies objects from multidimensional space to lower dimensional space and locates, analyze, categorize them, while retaining the original relationships between objects. The distance matrix is used as a measurement of the dissimilarity between stocks. After using SPSS (version: 19) multidimensional scaling (PROXSCAL) and setting spatial parameter as 3 dimensions, three dimensional coordinates of each stock can be acquired. Using these coordinates, the three dimensional spatial scattergram is drawn by Matlab as shown in Fig. 3. In order to facilitate the observation, a partial three dimensional spatial scattergram is drawn by using cluster 1, 9, 17, 18, 21 as an example, as shown in Fig. 4

.

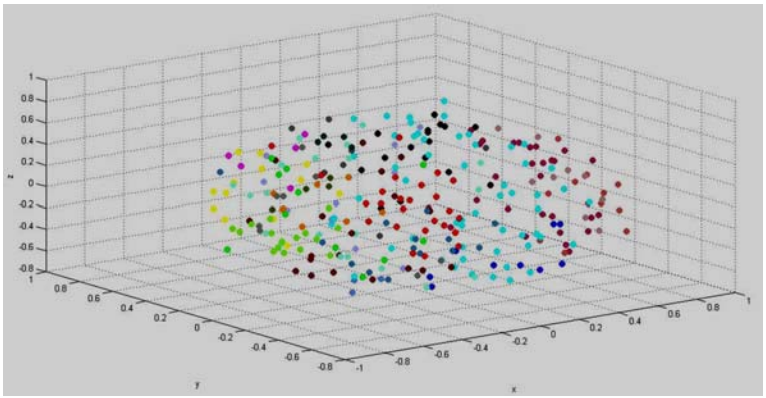

Fig.3 The Three Dimensional Spatial Scattergram of Stocks

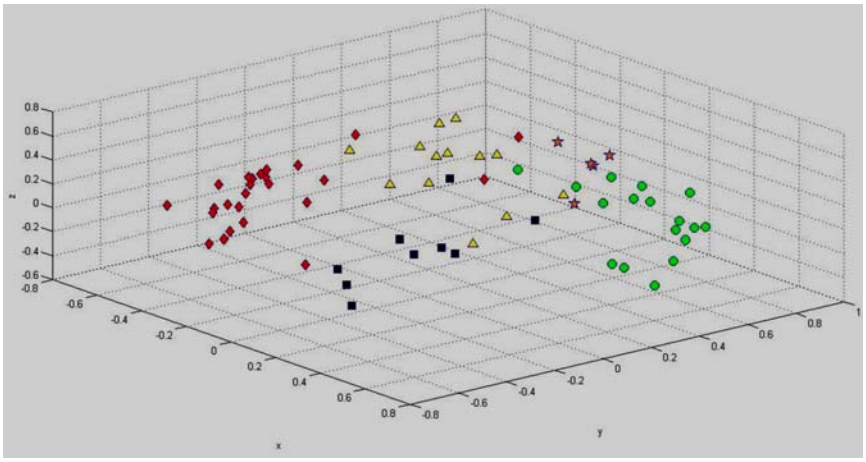

Fig.4 Partial Three Dimensional Spatial Scattergram (Cluster 1, 9, 17, 18, 21)

D. Case study

Xiaohang Zhang et al.[12] judged the cluster validity by the following guidelines: if in a certain cluster the majority of stocks belonged to the same industry, and that the small number of " outliers" were the result of misclassification, then they used error rates to determine the cluster quality. According to the classification results, almost the characteristics of every cluster are very distinct. There are 17 clusters without any outliers and the occurrence rate of 'outliers' is 7.36%. Thus we can recognize that the cluster result is satisfactory, and the measurement of correlation between stocks is relatively accurate.

Furthermore, we analyze the stocks in the same cluster, using cluster 21 as an example. In cluster 21, except for the stocks of securities companies, those 3 outliers is Jilin Aodong, a pharmaceutical company, Youngor, a garment company, and Liaoning Chengda a trade and pharmaceutical company. As they are divided into the cluster of securities companies, this seems to be a misclassified.

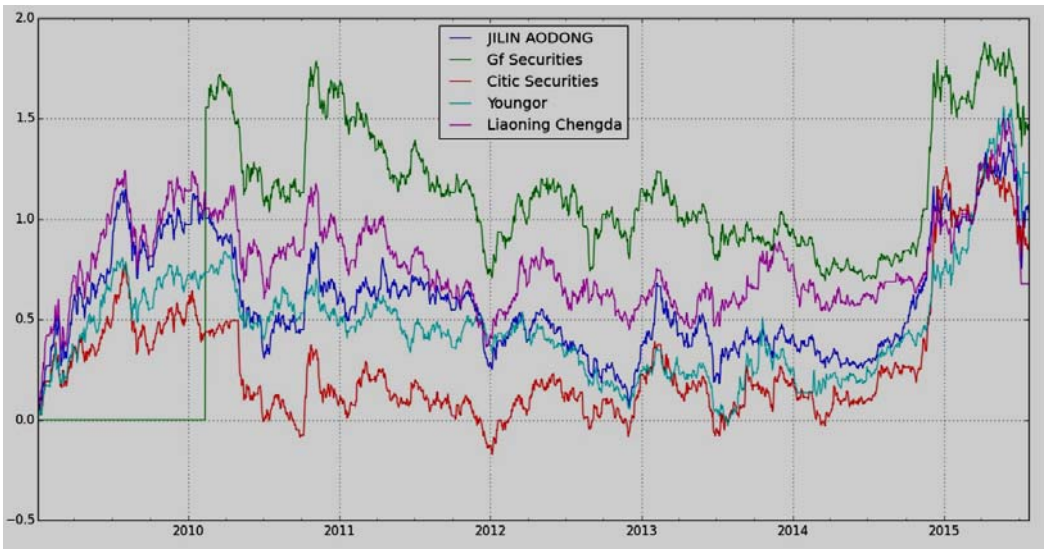

Fig.5 The Cumulative Log-return rate of Jilin Aodong, Youngor, Liaoning Chengda, Gf Securities and Citic Securities.

Nevertheless, according to the tendency of their cumulative log-return rate, shown as Fig. 5, their tendency is obviously similar. After doing some information collection, we find that Jilin Aodong and Liaoning Chengda are two of the major shareholders of Gf Securities, and Youngor is one of the major shareholders of Citic Securities. Therefore, that the tendency of Jilin Aodong, Youngor and Liaoning Chengda is correlative with securities companies

is reasonable.

E. ANOVA of time series

To analyze the cluster quality further, classification can be regarded as a factor and the different clusters are the distinct levels under that factor. We assume that the daily log-returns have significant difference under those different levels. Therefore, we use One-way ANOVA to analyze the log return rates grouped by clustering results every single day and count the number of days which have significant differences between groups under the significance level of 5%. The result is shown in Table IV.

TABLE IV. SUMMARY OF ONE-WAY ANOVA RESULTS

Significance level α	The number of days having significant differences between groups	Total number of days	Proportion
0.05	1526	1593	95.79%

Table IV shows that the log return rates which are grouped by clustering results have significant differences between groups in 95.79% days. On the one hand, because the number of clusters is chosen subjectively, it is OK only if the clustering result is a satisfactory solution. Apparently, that 95.79% days have significant differences is satisfactory. On the other hand, hierarchical clustering only uses the distances (correlation coefficients) of stocks which are the overall relationship between time series. However, the ANOVA has used information of each cross-section and this method confirms the reasonableness of clustering in other way.

IV.CONCLUSIONS

This paper has studied the correlation measurement method of clustering non-stationary time series. We use log-return series as analysis object, Spearman correlation coefficient between series as the measurement of correlation. According to this method, we cluster the stocks in CSI 300 and use three evaluation methods to observe the clustering effect. The results show that this cluster method can come up with a satisfactory classification results and the correlation between stocks is measured accurately at the same time. We also find that some clusters are not characterized by certain industries, such as 'Concept of Internet+ Stocks', and the deeper relationship between the stocks in it need to be further analyzed. On the other hand, when we use different period of time series for clustering, the cluster result will change. The reason and the market discipline that the changed result imply also need to be further studied in the future.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGMENT

This research was financially supported by the National Science Foundation (71371077) and Central Universities Basic Research Foundation (2014ZM0038).

REFERENCES

[1] Cherkassky V, Mulier F M. "Learning from data: concepts, theory, and methods." John Wiley & Sons, 2007.
[2] Plerou V, Gopikrishnan P, Rosenow B, et al. "A Random Matrix Approach to Cross-Correlations in Financial Data." Papers, 65(6):104-130, 2001.
[3] Fielitz B D. "Stationarity of Random Data: Some Implications for the Distribution of Stock Price Changes." Journal of Financial & Quantitative Analysis, 6(3):1025-1034, 1971.
[4] Xiaojun Zhao. "The Correlation and Complexity Analysis of Time Series." Beijing Jiaotong University. (in Chinese), 2015.
[5] Tsay R S. "Analysis of financial time series." John Wiley & Sons, 2005.
[6] Wu T, Yoshida S, Akiyama Y, et al. "Preliminary breakdown of intracloud lightning: Initiation altitude, propagation speed, pulse train characteristics, and step length estimation." Journal of Geophysical Research Atmospheres, 120, 2015.
[7] Martinez J, Martin J, Alonso A. "Teamwork competence and academic motivation in computer science engineering studies." Proceedings of Global Engineering Education Conference 2014 Ieee Global Engineering Education Conference 2014 Ieee 03 04 2014 05 04 2014 Istanbul Turkey, 2014.
[8] Lusardi A, Samek A, Kapteyn A. "Visual Tools and Narratives: New Ways to Improve Financial Literacy." Nber Working Papers, 2014.
[9] Pinacho R, Saia G, Meana J. "Transcription factor SP4 phosphorylation is altered in the postmortem cerebellum of bipolar disorder and schizophrenia subjects." European Neuropsychopharmacology, 2015.
[10] Lyu Y, Liang J, Yan J, et al. "On-line probabilistic dynamic security assessment considering large scale wind power penetration." Power System Technology (POWERCON), 2014 International Conference on. IEEE:2635-2641, 2014.
[11] Onnela J P, Saramäki J, Kaski K, et al. "Financial market-a network perspective"//Practical Fruits of Econophysics. Springer Tokyo: 302-306, 2006.
[12] Zhang X, Liu J, Du Y, et al. "A novel clustering method on time series data." Expert Systems with Applications, 38(9): 11891-11900, 2011.
[13] Mantegna R N. "Hierarchical structure in financial markets." The European Physical Journal B-Condensed Matter and Complex Systems, 11(1): 193-197, 1999.