A new data mining method improvement of time series based on auto regressive moving average model

Yingying Min

School of Engineering,Northeast Agricultural University,Heilongjiang 150030,Harbin,China
College of Computer Information Engineering,Harbin University of Commerce,Heilongjiang 150028, Harbin, China

Changlin Ao*

School of Engineering,Northeast Agricultural University,Heilongjiang 150030,Harbin,China
**Corresponding author, Email:myy80@126.com*

Abstract
The research technique of time series can be applied in many fields, and suitable time series model is a reflection of the characteristics of series. As the data mining method of time series based on the model can discover the internal laws of series, this method has a good research prospect. ARIMA (Auto Regressive Moving Average) model is a very important data mining model of time series, but this model often just studies a certain point in time, while, in fact, impact on the future often is the result from the work over a period of time. Therefore, on the basis of the past ARIMA model, this study uses regression theory to calculate the influence of a time period by the node, to improve the original model, and also predicts stock prices in the field of IT in the United States. And the experimental results show that the predictions of the improved model are more accurate and achieve good results.
Key words: ARIMA MODEL,DATA MINING,TIME SERIES,PREDICT

1. Introduction

The rapid development and extensive application of information technology enable enterprises, government departments and various other forms of organizations to accumulate a lot of data[1]. Past simple query and statistical techniques can only perform basic processing of data, but can not conduct a higher level of analysis, to automatically and intelligently transform the data to be processed into useful knowledge. Data mining gains extensive attention and is studied in depth in this context[2], and becomes an important research area which has made significant progress.

Data mining is a process to extract the knowledge which is implicit, unknown to people and has a potential value from the data. Data mining is known as one of the key technologies of future information processing[3]. In principle, data mining can be applied to any type of information sources[4], which include relation database, data warehouses, transaction database, other advanced database systems, flat files and data. Among these data sets, there exists a time relationship among the data of a type of data set, and such data are called time series. In the course of conducting data mining on time series, the time relationship must be considered among the data of data set, and

this type of data mining is called time series data mining[5]. Keogh[6]believes that time series is universal, and image data, text data, image data, handwriting data, brain scan data, etc. can all be seen as time series.

The research on how to effectively extract the potential useful knowledge from these vast amounts of complex time series has an important theoretical and practical significance. Therefore, TSDM has become an important branch of data mining research. TSDM research technique can be applied in many fields[7], and suitable time series model is a reflection of the characteristics of series. As the data mining method of time series based on model can discover the internal laws of series, this method has a good research prospect. The current time series models are mainly hidden Markov model, half hidden Markov model, BOX-Jenkins regression, ARIMA models and so on.

2. ARIMA model

ARIMA (Auto Regressive Moving Average) model is proposed by Box and Jenkins et al., which is used to conduct modeling on stationary time series. It is extensively used in economic and financial fields, and is typically used for prediction. ARIMA model is the integration of AR model and MA model, and it describes the system memory of its past state and system memory on noise when it enters system in the past. Based on stochastic process theory and mathematical statistics, this method studies the statistical law that random data series follows.

The general formula of ARIMA model (p,d,q) can be written as the following form:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + e_t - (\theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q}) \quad (1)$$

The model consists of two parts, the first part is the regression equation of p order, and the latter part is the error moving average (multiterm and form) equation of q order. This model reflects the prediction thinking of conducting q-order modification on the error of p-order regression model. Because this model appears in multiterm and form, p, q can stretch out and draw back according to the actual situation, so that the model can adapt to various types of time series.

3. Identify time series and choose the appropriate model

Time series can be divided into stationary and non-stationary categories, respectively applying ARIMA (p,q) and ARIMA (p,d,q) models, and the latter transforms the non-stationary model into stationary model by differential treatment[8]. When the non-stationary model includes seasonal factor, ARIMA (p,d,q) $(P,D,Q)s$ model is applied.

Therefore, the type of time series must be identified firstly, and then the right model can be found. The following steps are the process of category identification.

Step 1 Drawing dynamic discounting circle in the rectangular coordinate system can help people to have a more intuitive understanding on the trend of time series. This step allows you to roughly know whether the time series contains only random variation, or contains long-term trend, seasonal variation and so on.

Step 2 Calculating the autocorrelation coefficient respectively. When there is no significant difference between autocorrelation coefficient obtained and zero, the time series is stationary, and we can find the right model according to the steps. If the autocorrelation coefficient obtained is always significantly different from zero, the time series is not stationary.

Step 3 When that the time series is not stationary is determined, differential treatment is conducted on the time series, and the step is repeated, to determine whether the time series is stationary. The time series generally tends to be stationary after the second differential treatment.

Step 4 After the time series is determined as stationary, autocorrelation coefficient map and partial autocorrelation coefficient map can be applied to search the right ARIMA model, that is, to determine the appropriate p,d,q value. If the time series is seasonal, it can be observed from the autocorrelation function diagram that the autocorrelation coefficient which is significantly not zero at this time also appears periodically.

Step 5 Estimation of model parameters selects matrix estimation method, nonlinear least square method and inverse function method, which can estimate parameter value and fit ARIMA model of selected p,d,q combination.

4. Improvement on ARIMA model

In traditional ARIMA model, just some certain points in time are studied, but sometimes impact on the future is not only a point[9]. The final result change is caused by the accumulation of a period of time. So we try to improve ARIMA model from the aspect of a period of time. The specific method is to use equidistant node formula

$$\int_a^b f(x)dx \approx (b-a)\sum_{i=0}^{n} C_i^{(n)} f\left(x_i\right)$$

[10], and take n = 1, then:

$$\int_{t_1}^{t_2} T_t dt \approx \frac{1}{2}(T_{t-\tau_1} + T_{t-\tau_1-1}), t_2 = t-\tau_1, t_1 = t-\tau_1-1, \quad T = y_{t-1} \tag{2}$$

$$\int_{t_3}^{t_4} T_t dt \approx \frac{1}{2}(T_{t-\tau_2} + T_{t-\tau_2-1}), t_4 = t-\tau_2, t_3 = t-\tau_2-1, \quad T = y_{t-2}..... \tag{3}$$

$$\int_{t_{p+1}}^{t_{p+2}} T_t dt \approx \frac{1}{2}(T_{t-\tau_p} + T_{t-\tau_p-1}), t_{p+2} = t-\tau_p, t_{p+1} = t-\tau_p-1, \quad T = y_{t-p} \tag{4}$$

Thus equation (1) becomes

$$y_t = \frac{1}{2}\left[\phi_1(T_{t-\tau_1} + T_{t-\tau_1-1}) + \phi_2(T_{t-\tau_2} + T_{t-\tau_2-1}) + ... + \phi_p(T_{t-\tau_p} + T_{t-\tau_p-1})\right] + e_t - (\theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q}) \tag{5}$$

Using this model will give a more accurate prediction value.

The improved ARIMA model can be described by the following ways.

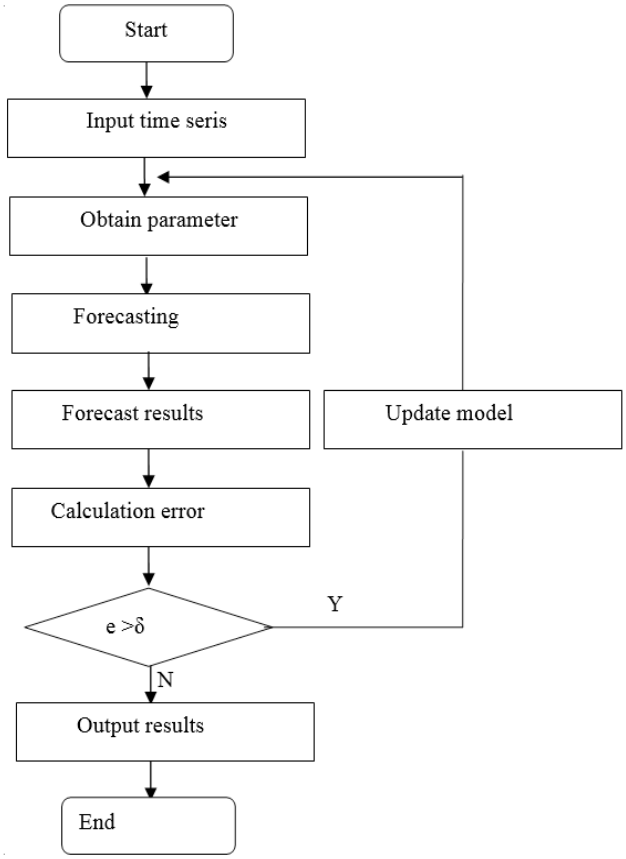

Figure 1. Flowchart of applying ARIMA model

5. Conducting prediction demonstration on stock price by applying ARIMA model before and after improvement

Prediction is to predict the stock price of the next day by the prediction on the yield rate of the next day and the current stock price[11], Java language is used to implement the relevant algorithms, and the operating environment is window XP[12]. The actual stock price series uses the stock price data set of US IT sector. Data from February 10, 2010 to September 10, 2011 are selected to conduct the test, and the data from September 13, 2011 to October 1, 2012 are applied for verification.

Because the object studied is ARIMA model, which is also a time series, so stationarity analysis should be first conducted on time series. There are numerous analysis methods on the stationarity of time series, and the method of correlation diagram is applied to conduct determination and study. The data from February 10, 2010 to September 10, 2011 are applied in correlation diagram and partial correlation diagram, and the autocorrelation coefficient is

$$r_k = \frac{\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \tag{6}$$

in which y_t is observed value, $\bar{y}$ is sample mean, T is sample size, and k is the lagging periods of lagging differential term. Here, $k = 16$ .In the correlation diagram, $AC = r_1, r_2, ..., r_k$, the partial autocorrelation coefficient of lagging k period is ϕ_{kk} , thus the partial autocorrelation coefficient is $PAC = \phi_{11}, \phi_{22}, ..., \phi_{kk}$,the calculation formula of Q statistics of lagging k period is

$$Q = T(T+2)\sum_{j=1}^{k}\frac{r_j^2}{T-j} \tag{7}$$

The correlation diagram and partial correlation diagram are presented as follows.

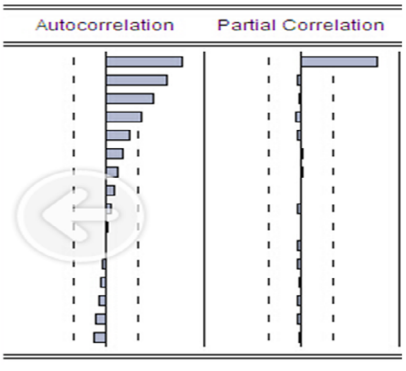

Figure 2. Correlation and partial correlation diagram

It can be observed from the correlation diagram and partial correlation diagram that the series is stationary. In order to further validate, feature root test is conducted on ARIMA model, model parameters pass t-test, and the reciprocal of eigenvalue is 0.92, thus satisfying the requirement of stationarity, which is shown in Table 1.

Table 1. Reciprocal test of eigenvalue

Vanriable	Coefficient	Std.Error	t-Statistic	Prob
C ARIMA	2205.157 0.923118	597.4120 0.044688	3.691182 20.65707	0.0008 0.0000
R-squared	0.667074	-	-	-
Adjusted R-squared	0.656985	-	-	-
S.E.of regression	1837.784	-	-	-
Sum squared resid	1.11E+08	-	-	-
Log likelihood	-311.7042	-	-	-
Mean dependent var	2035.910	-	-	-
S.D.dependent var	3137.892	-	-	-
Akaike info criterion	17.92595	-	-	-
Schwaez criterion	18.01483	-	-	-
Hannan-Quinn criter	17.95663	-	-	-
IMR	0.92			

The series is stationary in terms of both correlation diagram and partial correlation diagram and reciprocal of eigenvalue, and DF test of unit root should be conducted for further confirmation. For time series y_t , autoregressive model can be used to test unit root, and the formula is $y_t = \alpha y_{t-1} + u_t$; and use DF unit statistic to test unit root,

$$DF = \frac{\hat{\alpha} - 1}{s(\hat{\alpha})} = \frac{\hat{\alpha} - 1}{\sqrt{\frac{1}{T-1}\sum_{t=2}^{T} \hat{u}_y{}^2} \Big/ \sqrt{\sum_{t=2}^{T} y_{t-1}^2}} \quad (8)$$

OLS estimation formula of $\hat{\alpha}$ is

$$\hat{\alpha} = \sum_{t=1}^{T} y_t y_{t-1} \Big/ \sum_{t=1}^{T} y_{t-1}^2 \quad (9)$$

When DF> threshold, y_t is non-stationary; DF <threshold, y_t is stationary. u_t is random item. Test is conducted with Eviews6.0, and the test results are listed in the following table.

Table 2. DF test results of unit root

Variable name	ADF statistics	1% critical value	5% critical value	10% critical value	Probability	Conclusion
Improved model	14.2415	-3.53841	-2.93142	-2.610143	1	Non stationary
D(Improved model)	1.301127	-3.63237	-2.95411	-2.613534	0.98181	Non stationary
D(Improved model,2)	-7.36213	-3.63423	-2.84311	-2.314434	0.0326	stationary

According to test results, the values of the test results of time series unit root before improvements are all greater than the threshold of 1%, 5% and 10%, and the significance probability is greater than 0.05, indi-

cating that unit root exists, time series is not stationary, the original assumption cannot be denied, and there is a time trend. The values of the unit root test results of D (model before improvements) and D (model before improvements.2) are all less than the threshold of 1%, 5% and 10% , and the significance probability is less than 0.05, indicating that unit root does not exist, time series is stationary, and the original assumption is denied. It has the trend of long-term equilibrium, and the last D (model before improvements.2) is stationary.

Modeling following first uses the existing 700 data model to construct model, 300 data are used to test the model, the error term is written as e, the parameter ϕ_i of the model is estimated , $i = 1,2,3$ is taken here, the delay time selects $\tau = 1,2,3$, $\tau = 2,1,3$, $\tau = 3,2,1$, and estimation and correlation analysis are conducted on the parameters of the model based on statistical regression theory. The parameter value of the model is finally determined, as is shown in Table 3.

Table 3. Estimation value of model parameters

Model number	Delay time τ_1	Delay time τ_2	Delay time τ_3	ϕ_1	ϕ_2	ϕ_3	Correlation coefficient R
1	2	2	2	0.016000	-0.012195	0.015244	0.9762
2	2	2	3	0.018000	-0.012219	0.012575	0.9763
3	2	2	1	0.021000	-0.019408	0.014300	0.9772
4	2	3	1	0.011000	-0.037412	0.004044	0.9761
5	2	3	2	0.044000	-0.032128	0.003012	0.9759
6	2	3	3	0.007000	-0.035246	0.014098	0.9758
7	3	3	1	0.019000	0.008154	0.011213	0.9759
8	2	1	1	0.003000	-0.050150	-0.003009	0.9769
9	3	1	1	0.010000	-0.008080	0.002020	0.9758
10	3	2	1	0.027000	0.024666	0.021582	0.9776
11	3	2	2	0.022000	0.035787	0.021472	0.9769
12	3	2	3	0.024000	0.031762	0.021516	0.9765
13	3	1	3	0.005000	-0.003015	0.001005	0.9749
14	3	1	2	0.001000	0.002002	0.001001	0.9749
15	1	1	2	0.008000	0.048387	0.008064	0.9757
16	1	2	2	0.025000	0.067692	0.024615	0.9784
17	1	2	3	0.028000	0.064748	0.024666	0.9789
18	1	2	1	0.030000	0.057732	0.024742	0.9788

All the above models have passed t test (α=0.05)

It can be seen from the above conclusion that the correlation of Group 17 model is the highest and is in line with the requirements, so Group 17 model is selected for study.

Improved ARIMA model is applied to predict the stock price of US IT sector from September 2011 to October 2011, the accuracy rate of prediction is 98% (relative error is no less than 5%), and the predicted value, true value and error list are given below, which can more clearly reflect the accuracy extent of model. It is shown in Table 4 below.

Table 4. Price prediction on a stock of US IT sector from September 1, 2011 to September 24, 2011 (Using daily average value)

Predictive value of improved model	Predictive value of non-improved model	Actual value	Relative error
52.98	53.58	53	-0.05
53.00	53.03	56	-0.16
51.32	50.29	60	-0.16
50.12	49.50	59	-0.15
51.32	50.19	59	-0.16
51.12	49.90	58	-0.11
50.13	49.88	56	-0.11
49.65	49.85	56	-0.09

49.98	49.82	55	-0.08
51.23	49.66	54	-0.05
51.25	50.24	53	-0.01
50.26	51.50	52	0.04
50.21	51.89	50	0.03
51.42	52.92	51	0.05
53.87	53.82	51	0.02
52.36	52.82	52	0.01
51.13	52.70	52	0.03
53.65	52.68	51	0.03
54.32	52.76	51	0.03
51.49	52.61	51	0.05
51.26	52.45	50	0.03
54.21	52.33	51	0.03
50.12	52.51	51	-0.01
50.28	50.72	51	0.03

The errors of ARIMA model before and after updating are compared, which is shown in the Figure 3 below.

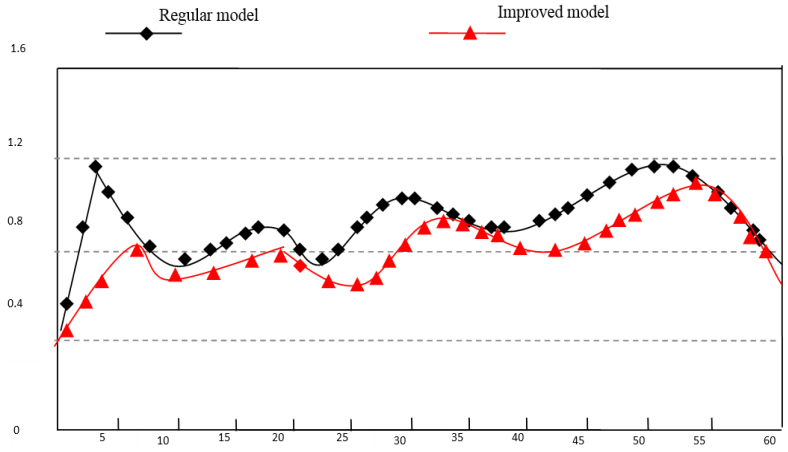

Figure 3. The error comparison

It is found that the error of improved ARIMA model is smaller, indicating that the improved ARIMA model is better. The predicted price and the actual price of stock are compared, which is shown in Figure 4.

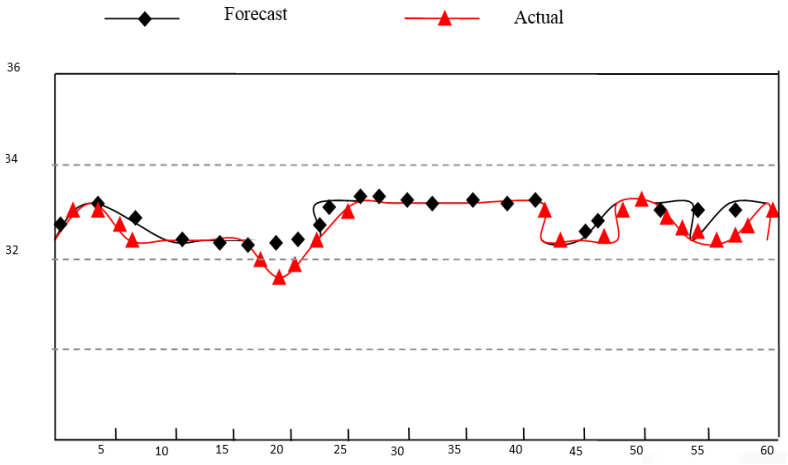

Figure 4. The comparison between forecast and actual stock price

The update model is applied to predict stock price, and the situation is shown in Figure 5.

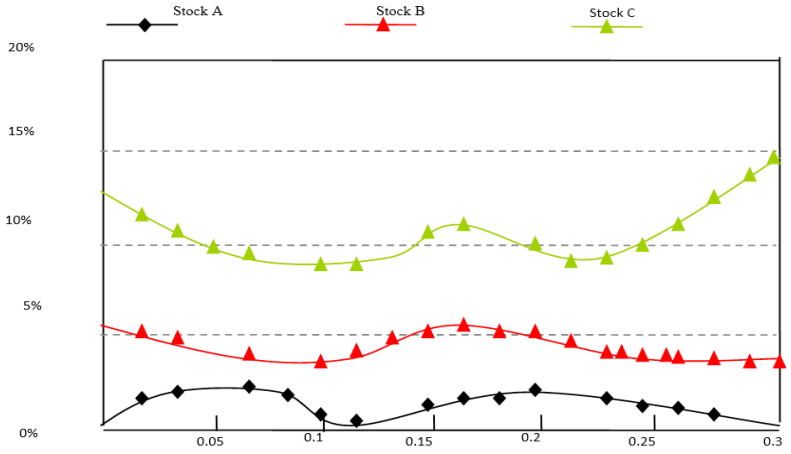

Figure 5. The price forecast of three stocks

As can be seen from the above graph, comparing with the ARIMA model before improvement, the error of ARIMA model after improvement is relatively smaller, and the stock prices predicted are more able to reflect the actual situation. It can achieve the basic forecast for the stock, and its forecasting results are better than that of the original ARIMA model.

Here the prediction of improved model is compared with that of other forecasting models, such as hybrid model weighted average method, ANN, original ARIMA and ARIMA model, and the data of several stocks from September 1, 2012 to September 30, 2012 are applied to predict the price. The comparative results of prediction accuracy are shown in Table 5.

Table 5. The comparison results of prediction accuracy of six stocks

The name of the stock	The original ARIMA	ANN	The mixed model of weighted average method	The improved ARIMA
Tsinghua Tongfang	1.803	2.746	3.234	1.659
Suning Appliance	1.324	1.998	1.857	1.286
Hainan Airlines	1.641	1.794	1.295	1.324
Founder Technology	1.189	0.86	1.097	0.731
Hundsun	0.986	0.586	2.978	0.512
China Southern Airlines	3.397	1.324	1.875	1.2286

It can be seen from the above chart that the prediction accuracy of the improved ARIMA model on the stock is the highest, which indicates that the improved ARIMA model has achieved the goal of improving prediction accuracy.

Then, the validity of the improved model prediction is tested, and compared with that of hybrid model weighted average method, ANN, original ARIMA and ARIMA model, the performance indicator applied is MAPE, and the MAPE performance indicator of the specific four models is shown in Table 6.

It can be seen from the above chart that the validity of the improved ARIMA model on the stock is the highest, which indicates that the improved ARIMA model has achieved the goal of improving validity.

Table 6. The performance index of MAPE of six stocks

The name of the stock	The original ARIMA	ANN	The hybrid model weighted average interpolation method	The improved ARIMA
Tsinghua Tongfang	1.903	2.646	2.234	1.759
Suning Appliance	1.314	1.898	1.557	1.276
Hainan Airlines	1.541	1.654	1.235	1.224
Founder Technology	1.089	0.816	1.087	0.631
Hundsun	0.965	0.486	2.778	0.412
China Southern Airlines	3.323	1.314	1.775	1.176

6. Conclusions

Based on ARIMA model, data mining model of time series, in order to better complete the prediction in a period of time, this paper makes improvement on ARIMA model, applies the improved ARIMA model to predict the stock price of US IT sector, and compares the error change of the ARIMA model before and after improvement. It finds that the error of the improved ARIMA model is smaller, compared with that of the original ARIMA model, and the stock prices predicted are more able to reflect the actual situation, thus achieving good results. However, ARIMA model cannot conduct long-term prediction because of its short-time characteristic, which remains to be further studied.

References

1. Han Wook-Shin,Lee Jinsoo,Pham Minh-Duc(2010)A Framework for Comparisons of Disk based on Graph Indexing Techniques. *Association for Computing Machinery*, 3(1), p.p. 449-459.
2. B. Phoophakdee, M J. Zaki(2007)Genome-scale Disk-based Suffix Tree Indexing. *Proc. Conf. on Management of Data*,New York, America,p.p.833-844.
3. Bishnu P.S, Bhattacherjee V.(2012) A dimension Reduction Technique for K-Means Clustering Algorithm.*Proc.Conf.on Recent Advances in Information Technology*,Piscataway, America,p.p. 531-535.
4. Jaziri Rakia,Benabdeslem Khalid,Elghazel Haytham(2010)A Graph based on Framework for Clustering and Characterization of SOM.*Proc. Conf.on Artificial Neural Networks*, Heidelberg, Germany, p.p.387-396.
5. CH IU S, KEOGH E, HART D, PAZZAN IM(2002). Iterative Deepening Dynamic Time Warping for Time Series.*Proc.Conf.on Data Mining*,Baltimore,America,p.p.148-156.
6. KEOGH E, CHAKRABARTI K, PAZZAN IM(2002)Locally Adap-tive Dimensionality Reduction for Indexing Large Time Series Databases. *ACM Transactions on Database Systems*, 27 (2), p.p.188-228.
7. Chang P C, Fan C Y, Lin J L(2011)Trend Discovery in Financial Time Series Data Using a Case Based Fuzzy Decision Tree. *Expert Systems with Applications*,38(5),p.p.6070-6080.
8. Sang T D, Thi T N, Woo D M.(2010)Standard Additive Fuzzy System for Stock Price Forecasting.*Intelligent Information and Database Systems*,59(9),p.p.279-288.
9. Hwang H, Oh J.(2010)Fuzzy Models for Predicting Time Series Stock Price Index.*International Journal of Control Automation and Systems*,8(3),p.p.702-706.
10. Gaweda A E, Zurada J M.(2003)Data-driven Linguistic Modeling Using Relational Fuzzy Rules.*IEEE Transactions on Fuzzy Systems*,11(12),p.p.121-134.
11. Bezdek J C.(1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic:Netherlands.
12. Bezdek J C, Ehrlick R, Full W.(1984)FCM: The Fuzzy C-means Clustering Algorithm. *Geoscience*,22(10),p.p.191-203.