Acta Technica *62* (2017), No. 4A, 721–728

Financial time series prediction research based on data mining

Jinhuan Wang[2], Yan L[2], Yulei Huang[2],Qing Lin[2]

Abstract. With the organic integration of computer science, artificial intelligence and mathematical statistics, data mining technology has been developed by leaps and bounds, which greatly promoted the development of financial time series mining technology. Data mining is the process of extracting non-trivial, implicit, unknown, potentially valuable information from large-scale data. Time series analysis, as a branch of mathematical statistics, has been unable to be used to discover more complex and subtle laws of data since the 1960s. For the above reasons, this paper combines data mining and financial time series to study. In view of this problem, this paper takes the clustering algorithm in data mining as the starting point, digs out the potential knowledge in the financial time series and forms the model to predict, and makes a deep research on the financial time series. The experimental results show that the new improved density-based spatial clustering algorithm can cluster the variable density data sets and adapt them according to the characteristics and attributes of the data set after the initial parameters are given. Compared with the traditional DBSCAN Algorithm, initial point optimization and parameter adaptive density spatial clustering algorithm can improve the quality of clustering to a certain extent. Compared with the traditional processing methods, the SVR algorithm based on OS-DBSCAN and particle swarm optimization has achieved good accuracy in the regression forecasting experiment of stock price and financial index, which can improve the accuracy of the next day's stock increase forecast. The data mining algorithm has achieved some effect and practicability in the prediction of financial time series.

Key words. Data mining; financial time series; clustering analysis.

1. Introduction

With the continuous development of database technology and the extensive application of database management system, the amount of data in the database has increased dramatically. The surge of data has hidden much important information, and people want to be able to analyze it at a higher level to use these data. But what the current database system can do is to access the data in the database, but

[1]Acknowledgement-Foundation item: Peihua special project of Xi 'an social science planning fund (No. 17PH07)

[2]Workshop 1 - ZTE Telecommunications College , Xi 'an Peihua University, Xi 'an , China

it cannot find hidden data behind the knowledge, resulting in "data explosion but the lack of knowledge" phenomenon. The results of the study found that the hidden information contained in these data contains a lot of useful information, the more important information is about the overall characteristics of these data description and prediction of its development trend, the information generated in the decision-making process has an important reference value. [1] It will provide decision-makers with important information knowledge, resulting in irreversible benefits. However, due to the complexity of the data, it is difficult to find the data needed by the user in the process of manually processing the data, so that much useful information is still implicit in the data and cannot be discovered and utilized. So how to dig out the potential, valuable information from these large, disorganized, potentially disturbing data presents new challenges for human intelligence processing. [2, 3]

Data mining (DM), also known as the database of knowledge discovery (referred to from the database, referred to as KDD), refers to a large number of incomplete, no noise, fuzzy, random application of practical data. To extract the hidden process, which is not known in advance, but is the potential value of information and knowledge of the complex process one, that is, according to the intended business objectives, a large number of enterprise data to explore and analyze, revealing the implied business laws and further modeled the technical process. [5] It is a wide range of interdisciplinary disciplines, including artificial intelligence, machine learning, mathematical statistics, database, pattern recognition, rough set, decision analysis and other related technologies.

The purpose of association analysis is to find hidden networks in the database, thus providing the necessary support for certain decisions. Clustering (clustering). Clustering task is to classify similar things into a class, the difference between the larger things in different classes. The difference between clustering and classification is that clustering does not depend on the pre-determined group, and no training set is needed. Clustering is a prerequisite for conceptual description and deviation analysis. Clustering techniques include traditional pattern recognition methods and mathematical taxonomy. Description and visualization (description and visualization). Data visualization is a descriptive and effective means of data mining. Making a meaningful visualization is not an easy task, but a good picture is more effective than hundreds of thousands of rules of association, because people are accustomed to abstracting useful information from the visual experience.[8]

2. Experimental procedure

2.1. Establishing the differential equation model

Differential equation is used to describe the time series data (Figure 1). Set up a set of observation time series of $x_t$value is: $X = [x_{t_0}, x_{t_1}, \cdots, x_{t_m}]^T$, in which, t_0 is initial time, $t = t_0 + i \cdot \Delta t$, Δt is time interval, The differential equation of time series modeling problem is solving a n order ordinary differential equation initial

value problem, that is:

$$\begin{cases} x^{(n)}(t) = f\left(t, x(t), x(t), \cdots, x^{(n-1)}(t)\right) \\ x^{(i)}(t)|_{t=t_0} = x^{(i)} (i = 0, 1, 2, \cdots, n-1) \end{cases} \tag{1}$$

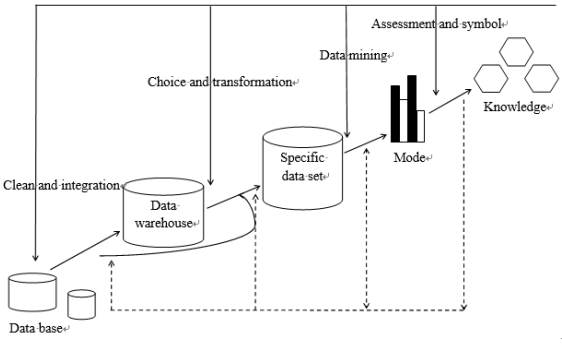

Fig. 1. The process of data mining.

The mean square error between the original time series and its solution $x^*(t)$ is:

$$\|x^*(t) - x(t)\| = \frac{1}{m}\sqrt{\sum (x^*(t_i) - x(t_m))} \tag{2}$$

We want it as small as possible.

For convenience, we consider a first order equation, the specific conditions are as follows:

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(t)|_{t=t_0} = x_i (i = 0, 1, 2, \cdots, n-1) \end{cases} \tag{3}$$

Using the central difference formula it can be obtained that: then we have the data matrix Y_1:

$$Y_1 = \begin{bmatrix} y_1(t_0) & y_2(t_0) \\ y_1(t_1) & y_2(t_1) \\ \vdots & \vdots \\ y_1(t_m) & y_2(t_m) \end{bmatrix} \tag{4}$$

Modeling problem can be converted into for a given data matrix Y_1, we want to seek for made in the form of differential equation:

$$\sqrt{\sum_{i=1}^{m} (y_1(t_i) - y_1^*(t_i))^2 + \sum_{i=1}^{m} (y_2(t_i) - y_2^*(t_i))^2} \tag{5}$$

Generalized to higher order form, for the original problem, there is an equivalent

of first order ordinary differential equations:

$$\begin{cases} x^{'}(t) = f(t, x(t)) \\ x^{(n)}(t) = \frac{dx^{(n-1)}}{dt} \\ x(t)|_{t=t_0} = x_i \, (i = 0, 1, 2, \cdots, n-1) \end{cases} \tag{6}$$

Using the central difference formula:

$$Y = \begin{bmatrix} y_1(t_0) & y_2(t_0) & \cdots & y_{n-1}(t_0) \\ y_1(t_1) & y_2(t_1) & \cdots & y_{n-1}(t_1) \\ \vdots & \vdots & \ddots & \vdots \\ y_1(t_m) & y_2(t_m) & \cdots & y_{n-1}(t_m) \end{bmatrix} \tag{7}$$

The original problem is converted into seeking a particular sequence of solution of differential equation, for a given data matrix Y, there is:

$$\min \sqrt{\sum_{k=1}^{n-1} \sum_{i=1}^{m} \left(y_1(t_i) - y_k^*(t_i)\right)^2} \tag{8}$$

2.2. Markov chain modeling

Suppose we have a random process $\{x_t, t \in T\}$, in which time is $T = \{0, 1, \cdots\}$, sample space is $I = \{a_1, a_2, \cdots, a_N\}$, if to any time t, and any sample a_i, we have:

$$P\left(x_t = a_i \left| x_{t-1} = a_{i_{t-1}}, x_{t-2} = a_{i_{t-2}}, x_1 = a_{i_1}\right.\right) = P\left(x_t = a_{i_1} \left| x_{t-1} = a_{i_{t-1}}\right.\right) \tag{9}$$

$\{x_t, t \in T\}$ is called Markov chain.
Its transformation probability is defined as:

$$p_{ij}(m, n) = P(X_n = a_j \,|x_m = a_i) = P(x_n = j \,|x_m = i) \qquad i, j \in S \tag{10}$$

In which,

$$p_{ij}(m, n) \geq 0 \quad i, j \in S \tag{11}$$

$$\sum_{j \in S} P_{ij}(m, n) = 1 \quad i \in S \tag{12}$$

So, the step k transformation probability is:

$$p_{ij}^{(k)}(m) = P(x_{m+k} = j \,|x_m = i) \quad i, j \in S \tag{13}$$

Normally, we have:

$$p_{ij}^{(0)}(m) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{14}$$

So, the step k transformation probability is:

$$P^{(k)} = \left\{ p_{ij}^{(k)}(m), \quad i, j \in S \right\} \tag{15}$$

With ergo Markov chain and stability. Has the ergo Markov chain in any sequence starting from which state, after a long enough time system is in stable state probability of j in a stationary distribution. Has the starvation of Markov chains, after a long time, the sequence will tend to be such a state, it has nothing to do with the initial state, its status in the $n+1$ and the former state of n period equal to the probability.

3. Time series forecasting

3.1. Data preprocessing

In order to verify the predicted effect of hybrid - OS algorithm on the financial index, and compared with other algorithms, this paper adopted the Taiwan Stock market composite index of China Taiwan's weighted mean number. TWII data have a long history, in this paper, the interception from January 4, 2010 to February 22, 2016, six years of data as the experimental data set. Then six years of data set is divided into 5 groups, each group of data span for five years, and put each set of data before 1340 session data is used as the training data, nearly two months behind the 40 days of data as test data, specific intercept situation such as table 1.

Table 1 5 groups data sets

Data set number	Training data set	Testing data set
1	2010/01/05- 2015/05/04	2015/05/05- 2015/06/29
2	2010/03/04- 2015/06/29	2015/06/30- 2015/08/26
3	2010/04/27- 2015/08/26	2015/08/29- 2015/10/25
4	2010/06/19- 2015/10/25	2015/10/26- 2015/12/20
5	2010/08/09- 2015/12/20	2015/12/21- 2016/02/22

From the financial trading software download refers to the number of time series data format for (transaction date, the opening price, the highest and the lowest price, closing price):

$$d_i = (date_i, open_i, high_i, low_i, close_i) \tag{16}$$

In this paper, on the basis of the sliding window width is 4 processes the trans-

action data set, we modified data formats as follows:

$$d_i = (date_i, \left(\frac{close_{i-1}}{close_{i-2}} - 1\right) \times 100, \left(\frac{open_i}{close_{i-2}} - 1\right) \times 100, \left(\frac{high_i}{close_{i-2}} - 1\right) \times 100 \\ \left(\frac{low_i}{close_{i-2}} - 1\right) \times 100, \left(\frac{close_i}{close_{i-2}} - 1\right) \times 100, \left(\frac{close_{i+1}}{close_{i-2}} - 1\right) \times 100) \tag{17}$$

In which, $\left(\frac{close_{i+1}}{close_{i-2}} - 1\right) \times 100$ is the close price in the next day, we treat is as the forecasting value. After such processing data is with the ratio of the value, the numerical change range is not big, but the basic fixed, so there is no normalized to the standard for data $(-1, 1)$, which can keep more numerical scale, retain more trades, and the inner link of data.

3.2. Parameter selection

Hybrid - OS parameter selection to TWII data are as shown in table 2.

Table 2 Hybrid - OS parameter selection to TWII data

OS-DBSCAN	Number	ε-SVR	Number	PSO
k	2000	Maximum of C	100	Local search ability
α	0.96	Minimum of C	1.005	Global search ability
β	1	Maximum of γ	1000	The largest number of evolution
		Minimum of γ	0.005	Population data
		ε	0.005	Cross certification k

3.3. Performance evaluation indicators

We want to test the hybrid - OS algorithm on the financial index prediction effect, this article through to evaluate statistical metrics, Mean square Error (MSE), Mean absolute Error. The average absolute error expression is as follows:

$$MAE = \frac{\sum_{i=1}^{n} |A_i - F_i|}{n} \tag{18}$$

The A_i is the actual value of the data points i, F_i is the data points i predicted, n is the number of all data points. The smaller the three evaluation indexes, said the more precise forecast, the larger the explain accuracy is not high.

Table 3 Hybrid - OS performance evaluation index

Data set	MSE	MAE	MAPE(%)
1	4018	51.03	0.8374
2	6113	60.47	0.9596
3	4802	57.21	0.9461
4	3684	47.06	0.7752
5	7836	68.21	1.0384

We can clearly find that the hybrid - OS algorithm in MAPE measure performance is best. It shows that the hybrid - OS algorithm can forecast algorithm than before more close to the actual value, is the stronger prediction ability. In addition to MAE and MSE evaluation index, this paper algorithm is better than the other two same experimental results.

4. Conclusion

This paper, by using the data mining technology to conduct the thorough research to the financial time series forecasting problems, combined with the existing clustering algorithm and support vector machine (SVM) algorithm, the improved DBSCAN clustering algorithm, the proposed can change the density of the initial point of clustering optimization and parameter adaptive spatial density clustering algorithm, and the hybrid particle swarm optimization algorithm of support vector regression machine, puts forward a kind of unstable predict financial time series data of the hybrid algorithm. Through Matlab platform to verify its validity and feasibility of using large-scale financial index data and vast amounts of stock data for the case study.Based on the experimental results can draw the following conclusion:

References

[1] Chen. M. Y,Chen. B. T: *A hybrid fuzzy time series model based on granular computing for stock price forecasting.*Information Sciences *294* (2015), 227–241.

[2] Sun. B. Q,Guo. H ,Karimi. H. R: *Prediction of stock index futures prices based on fuzzy sets and multivariate fuzzy time series.*Neurocomputing *151* (2015),1528–1536.

[3] Dash. R,Dash. P. K,Bisoi. R: *A self adaptive differential harmony search based optimized extreme learning machine for financial time series prediction.*Swarm and Evolutionary Computation *19* (2014),25–42.

[4] Wei. L. Y: *A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting.*Applied Soft Computing*42* (2016),368–376.

[5] Chen. M. Y, Chen. B. T: *Online fuzzy time series analysis based on entropy discretization and a Fast Fourier Transform.*Applied Soft Computing *14* (2014),155–166.

[6] Babu. C. N,Reddy. B. E: *A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data.* Applied Soft Computing*23* (2014) 27–38.

[7] Si. J,Mukherjee. A,Liu. B: *Exploiting Topic based Twitter Sentiment for Stock Prediction.* ACL(2)*2013* (2013), No. 4, 24–29.

[8] Cheng. C. H,Wei. L. Y: *A novel time-series model based on empirical mode decomposition for forecasting TAIEX.*Economic Modelling*36* (2014), 136–141.

Received November 16, 2016