Application of Data Mining in the Financial Data Forecasting

Jie Wang [1] and Hong Wang[2]

[1] College of Teacher Education, Shanxi Normal University, Linfen, Shanxi 041004, P.R. China
wjlkt@163.com
[2] College of Mathematics and Computer Science,
Shanxi Normal University, Linfen, Shanxi 041004, P.R. China
wangh@sxnu.edu,cn

Abstract. A method of processing financial data based on wavelet transformation is presented. The data of the financial is essentially an unfixed time sequence. Based on the wavelet transform, the series obtained after decomposition contains information. Basically, the wavelet decomposition uses a pair of filters to decompose iteratively the original time series. It results in a hierarchy of new time series that are easier to model and predict. Regarded as a signal, the time sequence is decomposed into different frequency channels (as a filtering step) .These filters must satisfy some constraints such as causality, information lossless, etc. And reconstruction is used to analyze and forecast the time sequence. Examples show that the new method is more effective than the traditional AR model forecast in some aspects.

Keywords: Data mining, Wavelet transform, reconstruction.

1 Introduction

In time series analysis and prediction, there are many methods or technologies which using artificial intelligence, machines study, neural networks, statistical theory, wavelet transform, genetic algorithm, the database technology and the policymaking theory concept, understands and analyzes the complex relations which hidden in the massive data, has the ability of processing different type data. The data mining algorithms have the validity and the probability. The data mining result usually is useful, definite, and can interactively excavate knowledge in the different level from different information source.

The wavelet analysis is a rapid development domain in modern mathematics. At present, it is widely applied in the signal analyzing and imagery processing. But the actually which using the wavelet analysis to analysis and forecast the financial's data are not often. In practical work, forecasting the financial data is extremely important, it plays important role in company's localization and market decision-making. Essentially speaking, the corporate financial data, such as sales volume, profit and disbursement and so on all is one kind of time series. They have the same characteristic as the usual analysis signal, these data can be forecasted through the wavelet analyzing. Moreover, This kind of corporate financial data also has its characteristic, if one kind of time series is steady, then may defer to the traditional method which using

D.-S. Huang et al. (Eds.): ICIC 2008, LNCS 5226, pp. 954–961, 2008.

model such as AR, MA or AR2, MA to forecast usually can achieve the quite good effect. But generally speaking, the corporate financial data stochastic undulation is very big, such as in this article sales volume achievement, the stability is very bad, it belongs to the model a non- steady sequence kind, this time these traditional forecast method effects are not very good Therefore, this article introduces the wavelet analysis method may decompose the signal wavelet to the different frequency channel. Because decomposes after the signal to be more unitary than on the frequency component the primary signal, and the wavelet decomposed to the signal has made smooth processing, so after decomposition, the time series stability is better than primitive time. After carrying wavelet decomposition on the some non- steady time series, it may be treat as the steady time series to process in the approximate significance, like this can use some traditional the forecast method to decompose after the time series to carry on the forecast, Moreover through the example proved this kind of forecast method effect is good.

Financial data analysis data mining system establish in some correlation rationales, like financial knowledge flexible expression, financial information resources appraisal, estimate and optimize information resources choice and so on. This is also the content, which is researched by precisely the modern finance theory. The financial data analysis data mining technology applies to the following domain: (1) Risk appraisal. The insurance is one risk service; the risk appraisal is one important work of insurance company. (2) Financial investment. The typical financial analysis domain has the investment appraisal and the stock transaction market forecasts, which is concerned with a matter development forecast. (3) Cheat recognition. The bank frequently has the cheating behavior, like the malignancy overdraws which brings the heavy loss to the bank.

2 The Wavelet Transform Theory

Usually, the wavelet decomposition and the heavy construction may realize through the Mallat algorithm. Supposes $\{\psi_{j,n,n\in z}\}\{V_j\}_{j\in z}$ is $L^2(R)$ center more than criteria analyzes φ is the criterion function, $\{\psi_{j,n,n\in z}\}$ is the wavelet base, Then influential the solution through the Mallat algorithm:

$$\begin{cases} c_k^{j+1} = \sum_{l\in Z} c_l^j \left\langle \varphi_{j+1,k}(x), \varphi_{j,l}(x) \right\rangle \\ d_k^{j+1} = \sum_{l\in Z} c_l^j \left\langle \psi_{j+1,k}(x), \varphi_{j,l}(x) \right\rangle \end{cases} \tag{1}$$

The heavy construction type is：

$$c_k^j = \sum_{l\in Z} c_l^{j+1} \left\langle \varphi_{j,k}(x), \varphi_{j+1,l}(x) \right\rangle + \sum_{l\in Z} d_l^{j+1} \left\langle \varphi_{j,k}(x), \psi_{j+1,l}(x) \right\rangle \tag{2}$$

Simply written as

$$c_k^{j+1} = \sum_l h_{l-2k} c_l^j, \qquad d_k^{j+1} = \sum_l g_{l-2k} c_l^j$$

$$c_k^{j+1} = \sum_l h_{k-2l}^* c_l^{j+1} + \sum_l g_{k-2l}^* d_l^{j+1}$$

Among them, $\{h_k\}_{k\in z} \in l^2(Z)$ is produced by type $\frac{1}{\sqrt{2}}\varphi\left[\frac{x}{2}\right] = \sum_k h_k^{\varphi}(x-k)$, may regard as the low pass filter coefficient; $\{g_k\}_{k\in z} \in l^2(Z)$ is produce by type $g_k = (-1)^k h_{1-k}^{\varphi}$, May regard as high passes the filter coefficient, c_k^{j+1} is the $\frac{1}{\sqrt{2}}\varphi\left[\frac{x}{2}\right] = \sum_k h_k^{\varphi}(x-k)\, c_l^i$ approximate signal, d_k^{j+1} is the c_l^j detail signal.

Table 1 listed the decision wavelet function quality essential property and some commonly used wavelet function nature.

Table 1. Commonly wavelet function's some nature

Wavelet function	Orthogonal	Tight structure	Structure length	Symmetry
Haar	have	have	1	yes
Daubechies	have	Approximate	2N-1	yes
Symlets	have	Approximate	2N-1	yes
Meyer	have	Limited length	no	yes

3 Time Series Analysis and Prediction

Let $x_1, x_2, ..., x_l$ be a measured time series. The objective is to predict the value of x_{k+p} using all the observations until the instant *k*. For this purpose, a functional relationship that maps the vector $[x_1, x_2, ..., x_l]$ and the value x_{k+p} is to be constructed with the principal concern of minimizing the prediction error. The optimal prediction sequence $x'_{p+1}, x'_{p+2}, ...$ minimizes the expectation (or generalization risk)

$$C = \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} E\{(x'_{p+1} - x_{p+1})^2 \,|\, x_k, x_{k-1}, ...\} \tag{3}$$

If the time series is random. It is given by

$$x'_{k+p} = E\{x_{k+p} \,|\, x_k, x_{k-1}, ...\} \tag{4}$$

Which cannot be computed since the information about the time series is limited to l measurements. The criterion to be minimized is the following empirical risk:

$$C_{emp} = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_i^{'})^2 \tag{5}$$

All the expectations are omitted if the time series is deterministic. The relationship between $x_{k+p}^{'}$ and the sequence $x_k, x_{k+1}, \ldots$ are supposed to be nonlinear of an unknown nature.

On the other hand, two problems arise. The model's order choice: When the order is small, the training is simple but the information (about the past) is not enough to predict accurately. However, when the order is high, the training is hard and the approximation error decreases slowly. This is the curse of dimensionality.

According to wavelet transform theory, signal $x(n)$'s wavelet series reconstruction can be replaced by the following finite sum.

$$x(n) = \sum_{j=0}^{J-1}\sum_{k\in Z} c_{j,k}\psi_{j,k}(n) \tag{6}$$

Formula (6) means to project the input signal $x(n)$ into corresponding scale's orthogonal subspace so as to recur signals in different distinguishing level.

To make $x_j(n) = \sum_{k\in Z} c_{j,k}\psi_{j,k}(n)$ is $x(n)$'s projective discrete form in wavelet subspace. The purpose of DWTAF is to produce the discrete reconstruction of $x_j(n)$ $(j = 0,1,\cdots J-1)$.

$v_j(n)$ (As is shown in fig.1) is the approximation of projection x_j:

$$v_j(n) = \sum_{k\in Z} \hat{c}_{j,k}\psi_{j,k}(n) \tag{7}$$

Where, $\hat{c}_{j,k}$ is wavelet coefficient $c_{j,k}$'s discrete approximation.

$$\hat{c}_{j,k} = \sum_{l} x(l)\overline{\psi}_{j,k}(l) \tag{8}$$

Formula (8) is put into formula (7), next result will be gained:

$$v_j(n) = \sum_{l} x(l)r_j(l,n) \tag{9}$$

Where $r_j(l,n) = \sum_{k\in Z}\overline{\psi}_{j,k}(l)\psi_{j,k}(n)$.

Formula (9) is $x_j(n)$'s approximation formula. It is the input signal $x(n)$ and filter $r_j(l,n)$'s discrete convolution form. These filters are consisted of wavelet $\psi(t)$'s inflation and convolution after sampling. They are bandpass filters with constant bandwidth/center frequency ratio to realize input signals' multi-distinguishing space reconstruction.

Under the supposition of time-steadiness and orthogonal, Formulas (10) and (11) can be gained:

$$r_j(l,n) = r_j(l-n) \quad , \quad r_j(m) = r_0(2^j m) \tag{10}$$

Then

$$v_j(n) = \sum_l x(l) r_j(l-n) \tag{11}$$

The discussion above is about the signal demonstration pf discrete wavelet transformed. Next the discrete wavelet transforms adaptive LSM algorithm is to be deduced.

If

$$\mathbf{V}(n) = [v_0(n), v_1(n), \cdots, v_{J-1}(n)]^T \tag{12}$$

$$\mathbf{x}(n) = [x(n), x(n-1), x(n-2), \cdots, x(n-N+1)]^T \tag{13}$$

$$[\mathbf{W}]_{jm} = r_j(m), \qquad j = 0,1,\cdots J-1, m = 0,1,\cdots N-1 \tag{14}$$

$$\mathbf{B}(n) = [b_0(n), b_1(n), \cdots, b_{J-1}(n)]^T \tag{15}$$

Then

$$\mathbf{V}(n) = \mathbf{W}\mathbf{x}(n) \tag{16}$$

Where $\mathbf{W}$ is the wavelet transform matrix, its dimension is $J \times N$. $\mathbf{V}(n)$ are the input signals after passing through discrete transform filter.

In the structure of DWTAF, filter $r_j(n)$ $(j = 0,1,\cdots J-1)$ and multi-group delay line coefficient $b_j(n)$ $(j = 0,1,\cdots J-1)$ constitute expected signal $d(n)$'s predictor. Such prediction is based on input $x(n)$'s J continuous linear combination. The adaptive filtering output signals are:

$$y(n) = \mathbf{V}^T(n)\mathbf{B}(n) = \sum_{j=0}^{J-1} v_j(n) b_j(n) = \sum_{j=0}^{J-1}\sum_{i=0}^{N-1} x(n-i) r_j(i) b_j(n) = \sum_{i=0}^{N-1} \alpha_i x(n-i) \tag{17}$$

Where, $\alpha_i = \sum_{j=0}^{J-1} r_j(i) b_j(n)$.

The adaptive filtering error signals are:

$$e(n) = d(n) - y(n) \tag{18}$$

Coefficient update formula is:

$$\mathbf{B}(n+1) = \mathbf{B}(n) + 2\mu e(n)\mathbf{V}(n) \tag{19}$$

The algorithm convergence condition is

$$0 < \mu < \frac{1}{\lambda_{V\max}} \tag{20}$$

To make $\mathrm{R}_{\mathbf{VV}}$ is $\mathbf{V}(n)$'s self- correlation matrix, $\mathbf{R}_{\mathbf{V}d}$ is $\mathbf{V}(n)$ and expected signal $d(n)$'s cross correlation matrix. So $\lambda_{V\max}$ in formula (21) is $\mathrm{R}_{\mathbf{VV}}$'s maximum eigen value.

$$\mathbf{B}(n+1) = \mathbf{B}(n) + 2\mu e(n)\mathbf{D}_w^{-1}\mathbf{V}(n) \tag{21}$$

4 Using the Wavelet Analysis to the Financial Data for Forecasting

The main motivation of using the wavelet decomposition is the easy analysis of the obtained series. The corporate financial data can be regards as one kind of time series. It is non-steady time series, then namely $c^0 = [x_1, x_2, ..., x_n]$. Carries on the financial data using the wavelet before the forecast, must carry on wavelet processing first to c^0. The step of the forecasting method of the financial data is explained as following:

1) Selecting suitable wavelet function to carry on the decomposition. In actual forecast process, the data can be dealt with according to the different question choice and different wavelet mother function, at the same time unifies the consideration different wavelet mother function the different characteristic to forecast the value of the influence. Compares various wavelets function processing signal through the analysis, the result and compares with the theory result, determined the wavelet function quality with the error which produces in the processing.

2) Definition the suitable layer of wavelet decomposition. The greatest criterion determination will be advantageous in a group of forecast. The more criterions are big, the more computation workload is also bigger, and the error also can increase. But, the ruler goes past is greatly more advantageous to from the deeper level clear signal trend analysis, it can cause the time series to be steadier in the actual process. When the time series data's quantity which will be forecasted is not very big, the decomposition layer generally is the 3~5 level.

In fact, the trend c_{nk} contains the slowest change. It is practically noise free. The detail series d_{jk} contains the dynamics at a certain intermediate scale. The greater is *j*, the slower are the change. The low detail series may also be corrupted by noise. As a consequence of this property, training an estimator on these time series is simpler than on the original data. However, in low level detail series, the information is totally embedded in noise, one can simply put at zero the corresponding predictions.

3) Decomposing the time series data c^0.

Processing wavelet transform coefficients by formula (6) to (17). Then, the forecasting result is obtained.

5 Application

The series studied here represents the bank savings of a province in 2004, as figure 1. The performance of the forecasting is measured by formula (6) to (17). The wavelet function is 'db4'; the layer of wavelet decomposition is 3. The figure 2 is wavelet decomposition coefficients. The figure3 is result of financial data forecasting.

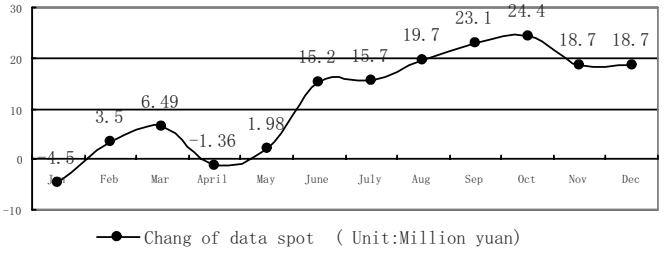

Fig. 1. The bank savings image of a province in 2004

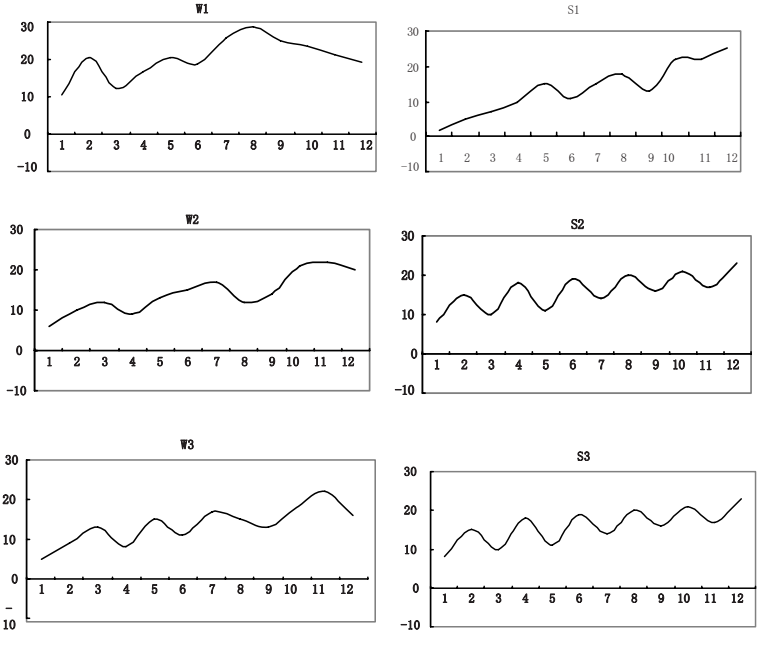

Fig. 2. The high frequency and low frequency of data composition

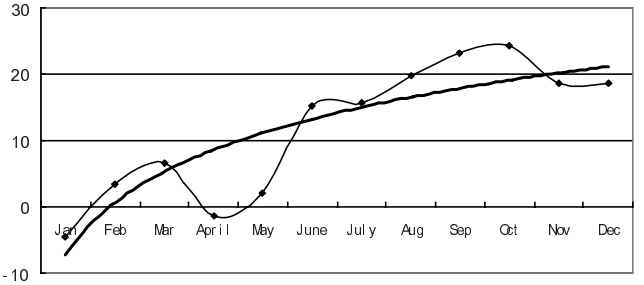

Fig. 3. The result of financial data forecasting (see the thick string)

6 Conclusion

Time series clustering has been shown effective in providing useful information in various domains. There seems to be an increased interest in financial time series clustering as part of the effort in temporal data mining research. It is based on the wavelet transform. The series obtained after decomposition contains information. It may also be used to separate noise from relevant information. Under some conditions, the wavelet decomposition method reduces the empirical risk. The results obtained through the bank saving of a province in 2004 prove the approach effective.

References

1. Casdagli, M.: Nonlinear Prediction of Chaotic Time Series. Physica D 35, 335–356 (1989)
2. Vieu, P.: Order Choice in Nonlinear Autoregressive Models, vol. 26, pp. 304–328 (1995)
3. Li, Z.C., Luo, J.S.: Wavelet Analysis and Its Application. Electronics industry publishing company, Beijing (2003)
4. Zhao, K., Wang, Z.H.: Wavelet Analysis and Its in Analytical Chemistry Application. Geological Press, Beijing (2000)
5. Soltani, S., Boichu, D., Simard, P., Canu, S.: The Long-Term Memory Prediction by Multiscale Decomposition. Signal Processing 80, 2195–2205 (2000)
6. Xu, K., Xu, J.W., Ban, X.J.: Based on Wavelet Certain Non- Steady Time Series Forecast Methods. Electronic journal 29, 566–568 (2001)