Double layer interval graph model: the universal tool for data driven market analysis and forecasting

Nikolay V. Markov*, Andrey V. Dmitriev, Svetlana V. Maltseva

National Research University Higher School of Economics, 33 Kirpichnaya st., Moscow, 105187, Russia

Abstract

This scientific work is dedicated to the development, improvement and application of double layer interval weighted graphs (DLIG) for non-stationary time series forecasting. This model appears to be the universal and easy-to-use tool for modeling the non-stationary time series and forecasting. We observe the double layer version of the model because it’s the most representative way in the sense of main idea though you can add several layers more for different purposes. The first layer of the graph is based on empirical fluctuations of system and displays the most potential fluctuations of the system at the time of system training. The second layer of the graph as a superstructure of the first layer displays the degree of modeling error and it’s connected with the first layer nodes by edges. The second layer is the way of supervised training implementation with the aim of error minimization.

Keywords: market analysis, machine learning, DLIG, non-linear dynamics, forecasting

1. Introduction

Today market specialists need the universal tool for market analysis especially in time series forecasting. The reason for this is the presence of numerous factors of probabilistic nature in time series, which are often not possible to account because of ignorance of the laws of random variables distribution or cumbersome computational procedures. On the other hand, the risk factor is a strong incentive for saving money and resources and optimization of all business processes.

Market risks are associated with fluctuations in prices for different kinds of goods, services, financial instruments, and more. In this regard, in the conditions of market crisis, characterized by sharp price...
fluctuations, the major problem is the choice of the mathematical model for the evaluation, analysis and prediction of non-stationary time series.

There are lots of statistical and structural models of analysis and forecasting of time series [1], however, due to the low efficiency of most of them in conditions of unstable markets there are some interesting and innovative approaches to forecasting such as, for example, models based on fuzzy logic and Markov chains [2], [3], [4]. The main purpose of both statistical and structural prediction models is the minimization of the prediction error [5]. We are interested in dynamic models (as a part of structural models) based on the differential equations. We are also interested in dynamic chaos modeling [6] and in adaptive models, that are capable, based on the training set, independently configure themselves to identify explicit and implicit “preconditions” to sharp fluctuations inside of the time series.

For example, let’s overview a mathematical model based on fuzzy logic. It involves the use of nonstandard methods of approximate calculations. In this model, the universal set, formed from the values of the time series, can be divided into intervals. The resulting intervals are fuzzy sets (A_i) with the certain logical relations for them. The disadvantage of this approach is the lack of well-defined, well-grounded algorithm for estimating the optimal number of intervals for a specific task. The user model defines a number of intervals divided oscillation amplitude values of the time series by himself. In addition, the length of intervals of the plurality of universal set affects the prediction accuracy [7], [8].

During the research of the model described below, it was decided to borrow some effective techniques of modern machine learning models, including neural networks and fuzzy logic models. Thus, for example, to break the universal set, we used the scale of “1-9”, proposed in [9]. The reasoned usage of scale “1-9” in the conditions of our problem helped us to break the universal set into 18 intervals; later (after the implementation of a software product) that gave us the acceptable ratio of CPU time used, as well as the accuracy of modeling and forecasting. It should be noticed that there are no clear recommendations on the conditions of division into intervals. T. Saaty applied the nine-level ranking to the hierarchy analysis process in decision-making, not for time series analysis. The next step in the modeling of fuzzy time series is the formation of fuzzy logical relationships (A_i → A_j) and their grouping scheme.

\[
\begin{align*}
A_i \rightarrow A_{j1} & \quad \Rightarrow \quad A_i \rightarrow A_{j1}, A_{j2} \\
A_i \rightarrow A_{j2} & \quad \Rightarrow \quad A_i \rightarrow A_{j1}, A_{j2}
\end{align*}
\] (1)

In continuation it should be said that the goal of this research is not only to design a new approach to time series forecasting, but also to design its practical implementation.

2. Algorithm of double layer interval graph training

The basis of the proposed prediction model is system training based on structural features of the time series (training set).

2.1. Primary training

Consider the non-stationary time series \{F(t_0), F(t_1), \ldots, F(t_n)\}. Nonstationarity of time series is due to the high financial indicators fluctuation amplitude on financial market, characterized by rapid changes and price shocks. We form the universal set \{ΔF_r(t_0), ΔF_r(t_1), \ldots, ΔF_r(t_n)\} from the set of relative variations of adjacent values of time series, that can be counted using formula:
\[ \Delta F_r(t_i + 1) = \frac{F(t_i + 1) - F(t_i)}{F(t_i)}, i = 0 \ldots n \] (2)

To split the time series fluctuation amplitude into intervals we find maximum relative negative and positive variations:

\[ \Delta F_{\text{max}}^- = \min \{ \Delta F_r(t_1), \Delta F_r(t_2), \ldots, \Delta F_r(t_n + 1) \}, \]
\[ \Delta F_{\text{max}}^+ = \max \{ \Delta F_r(t_1), \Delta F_r(t_2), \ldots, \Delta F_r(t_n + 1) \} \] (3)

Using scale of ",1-9" of T. Saaty for (3) we break the universal set into 18 intervals:

\[ A_1 = [\Delta F_{\text{max}}^-, \Delta F_{\text{max}}^- - d_1^-] , A_2 = [\Delta F_{\text{max}}^- - d_1^- , \Delta F_{\text{max}}^- - 2d_1^-] , \ldots \]
\[ A_9 = [\Delta F_{\text{max}}^- - 8d_1^- , 0], A_{10} = [0, \Delta F_{\text{max}}^+ - 8d_1^+], \ldots \]
\[ A_{17} = [\Delta F_{\text{max}}^+ - 2d_1^+, \Delta F_{\text{max}}^+ - d_1^+], A_{18} = [\Delta F_{\text{max}}^+ - d_1^+, \Delta F_{\text{max}}^+] \] (4)

In (4) the lengths of intervals:

\[ d_1^- = \frac{\Delta F_{\text{max}}^-}{9}, d_1^+ = \frac{\Delta F_{\text{max}}^+}{9} \] (5)

"9" in the denominator is directly related to the scale of ",1-9" T. Saaty, determining the degree of significant deviations from zero. In this sense, the values of "linguistic" variable: \( A_1 \) – maximal variation, \( A_2 \) - very big variation, \( A_3 \) - big variation, \( A_4 \) - variation above the mean, \( A_5 \) - mean variation, \( A_6 \) - variation below the mean, \( A_7 \) - small variation, \( A_8 \) - very small variation, \( A_9 \) - very small variation (symmetrically for \( A_{10} \) - \( A_{18} \)). These definitions of "linguistic" variable values are directly connected with the specificity of solving problem: identifying the risk degree and decisions making according to the model.

Then we sign each element \( \Delta F_r(t_i + 1), i = 0 \ldots n \) as a corresponding interval \( A_i \).

Define the structure of oriented graph \( G_A = \langle IV_A, E_A \rangle \), whose nodes are the obtained intervals:
\( IV_A = \{ A_1, A_2, \ldots, A_{18} \} \). The edges of the graph are defined by the logical relations:

\[ E_A = \left\{ A_i, A_j \right\} | A_i \rightarrow A_j \} \] (6)

Relations (6) describe the sequence of relative variations in the structure of time series: variation on step \( (k + 1) \), which is related to interval \( A_i \) follows the variation on step \( k \), which is related to interval \( A_i \). Respectively, we obtain the edge \( E_A = \left\{ A_i, A_j \right\} | A_i \rightarrow A_j \} \), which attests that interval \( A_j \) follows interval \( A_i \) at the \( (k + 1) \) step. There can be several ordered in this way sequences of pairs, so each of them will be called...
“transition”. With possible set of transitions from one interval node to another there can be only one corresponding edge in the graph.

For the definition of relationships between the adjacent values \( \Delta F_r(t_j) \) and simplification of model computation we shall use grouped relations \( R_j = \left\{ A_i, (A_{j_1},...,A_{j_n}) \right\} \rightarrow (A_{j_1},...,A_{j_n}) \).

Define the modeled values of time series using formula

\[
\hat{F}(t_{i+1}) = F(t_i) + F(t_i) \varphi(A_i)
\]  

(7)

In (7) \( A_i \) - interval at step \( t_i \), \( \varphi(A_i) \) - the value of modeling variation for the node \( A_i \).

\( \varphi(A_i) \) is defined by the formula:

\[
\varphi(A_i) = A_{mid} \times \mu(A_i)
\]  

(8)

In (8) \( \varphi(A_i) = (\overline{A}_1,...,\overline{A}_{18}) \) - vector of mean values of each interval, \( \mu(A_i) \) - values of each weight function for the node \( A_i \).

Weight function for each node in (8) is defined in the following way:

\[
\mu(A_i) = \begin{bmatrix}
\lambda_1 \\
\vdots \\
\lambda_{18}
\end{bmatrix},
\]

(9)

where \( \lambda_k = \frac{c_k}{N_i} \), \( c_k \) - the amount of transitions \( A_i \rightarrow A_k \), \( N_i \) - the amount of transitions from the central node \( A_i \).

2.2. Secondary training

The first and the second steps of the secondary training algorithm is similar to the corresponding ones from the primary training. The only difference is usage of the universal set \( \left\{ \Delta \overline{F}_r(t_1),...,\Delta \overline{F}_r(t_n) \right\} \) from the relative variations between the actual and modeled values of time series:

\[
\Delta F_r(t_i) = \frac{\hat{F}(t_i) - F(t_i)}{F(t_i)}, i = 0...n
\]  

(10)

We use the partition of the universal set into 18 intervals similarly to (4) \( \left\{ B_1,...,B_{18} \right\} \).

Define the superstructure (the second layer) for the graph \( G_A \), which consists of a set of nodes \( IV_B = \left\{ B_1,...,B_{18} \right\} \). In this case the relations between the nodes of the primary training graph \( G_A \) and the nodes of the graph \( G_B \) are the following:
\[ E_{AB} = \{ (E_A, B_j) \mid E_A \rightarrow B_j \} \]  

(11)

For the convenient computation lets use the grouped relations between the nodes of the layers:

\[ R_{AB} = \{ (E_A, (B_{j1}, \ldots, B_{jn})) \mid E_A \rightarrow B_{j1}, \ldots, B_{jn} \} \]  

(12)

With double “layerness” of interval graph in mind the modeled values of time series are defined using the formula:

\[ \tilde{F}(t_{i+1}) = F(t_i) + (\varphi(A_i) - \varphi(B_i)) \times F(t_i) \]  

(13)

where \( A_i \) - interval of relative variation at step \( t_i \), \( B_i \) - interval of error at step \( t_i \), \( \varphi(B_i) = B_{mid} \times \mu(B_i) \) - the value of modeled variation for the node \( B_i \), \( B_{mid} = \overline{B_1 \ldots B_{18}} \) - vector of mean values of each partition interval.

Weight function in (13) for each node is defined in the following way:

\[ \mu(B_i) = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_{18} \end{pmatrix} \]  

(14)

where \( \lambda_k = \frac{c_k}{N_i} \), \( c_k \) - the amount of transitions \( E_A \rightarrow B_k \), \( N_i \) - the amount of transitions from the central node \( E_A \).

Thus the proposed prediction model can be represented in a two layer view: a primary layer and secondary system training layer. A mathematical model of forecasting can be represented as a double layer interval weighted graph (DLIG) (Fig. 1).

The first layer is formed during the primary training and it is an oriented graph whose nodes are the intervals (interval nodes). Given graph displays all the possible fluctuations of the system in the interval of time in which a system was trained.

The second layer of the model is represented by a superstructure over the first layer graph. These relations show which intervals include the modeling errors after the primary system training. Thus we minimize the error of time series modeling (training set).
3. Algorithm of double layer interval graph training

To analyze the effectiveness and validation of this training model we have made 21-day forecast of the closing prices for gold using the daily accounting data from Bloomberg system. The training set includes the prices of gold above during the period from 01.05.2006 to 02.13.2015 (Fig. 2), that were converted to the set of variations between the adjacent values of the gold prices time series (Fig. 3). The practical significance of this prediction can be caused by various government or business purposes: from the analysis of the current economic situation to hedging processes.

Before the analysis of learning outcomes and forecasting system should describe the available data. Gold prices had an average volatility; from mid-2006 until mid-2007 values remain stable (no sharp variations); there is a relatively stable increase in prices values since early 2009 to mid-2011, where the maximum value. This is followed by a period of “intermittency”, accompanied by a significant local irregular values.
When testing the accuracy of the system of training and further forecasting, the model showed the following results: average prediction error without the second layer of minimization - 2.98% (single interval weighted graph without secondary education); using minimization of the second layer - 2.02% (two-layer interval weighted graph).

We used a standard type of model values error $\varepsilon$: 

![Graph showing time series of training set](image)

Fig. 2. Time series of training set (the period from 01.05.2006 to 02.13.2015)

![Graph showing deviations inside of training set](image)

Fig. 3. The deviations inside of the training set (the period from 01.05.2006 to 02.13.2015)
\[ \varepsilon = \frac{|\widetilde{F}_t - F_t|}{F_t} \]  \hspace{2cm} (15)

where \(\widetilde{F}_t\) is a simulated value, \(F_t\) actual value for the corresponding time \(t\).

The result of the 21-day TLIG forecast of behavior of prices of gold, as well as the results of the comparative characteristic prediction using neural network with the structure of 4:10:1 (the number of input nodes, the number of neurons in the hidden layer, the number of output nodes) is shown in Fig 4.

The input to the neural network was fed by first four values, the reference value was the fifth one, then we produced a shift to a single value, and training continued so on.

According to the test results it is obvious that TLIG-trained system is able to "control" the nonlinear time series trend in contrast to the neural network, which would correspond to the customer’s requirements in the risk assessment; the forecast accuracy is higher as well. The comparative analysis of forecast accuracy using TLIG and neural network is shown in Table 1.

![Fig. 4. The comparative analysis of forecast accuracy](image)

| Training model                      | Error, % |
|-------------------------------------|----------|
| DLIG, single layer                  | 2.98%    |
| DLIG, double layer                  | 2.02%    |
| Neural network (4 : 10 : 1)         | 4.71%    |

Table 1. Comparative analysis of forecast errors
4. Conclusions

Our research leaves several questions unanswered. We do not prove that our algorithm of Section 2 has better error than some neural networks in other formations and in other assumptions, but still this supervised training approach works and “controls” the non-linear behavior by catching the patterns inside of the training set and converting them into relations, so it can be implemented for solving different business problems related to the forecasting problem (e.g., risk hedging).

The potential of this approach is much wider than it’s shown here, but DLIG model is already successfully tested and proved its solvency in Novolipetsk Steel Metal Company for the forecasting of steel billet prices and market analysis in the year of 2014.

References

[1] Jingfei YM, Power system short-term load forecasting. Thesis for Ph. Degree. Germany, Darmstadt, Elektrotechnik und Informationstechnik der Technischen Universitat, p. 139.
[2] Song Q, Chissom BS, 2006. Fuzzy time series and its models. Fuzzy Sets and Systems, Vol. 54: p. 89.
[3] Chen SM, 1996. Forecasting enrollments based on fuzzy time series. Fuzzy Sets and Systems, Vol. 81: p. 311–319.
[4] Jilani TA, Burney SMA, C. Ardil C., 2007. “Fuzzy metric approach for fuzzy time series forecasting based on frequency density based partitioning”, Proceedings of World Academy of Science, Engineering and Technology, Vol. 23: p. 333–338.
[5] Brockwell PJ, Davis RA, 2002. Introduction to Time Series and Forecasting. Springer-Verlag, 2nd ed.
[6] Dmitriev AV, Maltseva SV, Markov NV, 2014. “Crisis forecasting in ask and bid prices formation systems in case of precious metals”, Global Interdisciplinary Business-Economics Advancement Conference. Conference Proceeding, p. 11–18
[7] Chen SM and Hsu CC, 2004. A new method to forecast enrollments using fuzzy time series. Int. Journal Applied Science and Engineering, Vol. 2, p. 234–244
[8] Huarng K, 2004. Effective lengths of intervals to improve forecasting in fuzzy time series. Fuzzy Sets and Syst., Vol. 123, p. 387–394
[9] Saaty TL. Decision Making for Leaders: The Analytical Hierarchy Process for Decisions in a Complex World, New Edition. RWS Publications 2012