An Implementation of Prediction Calculation Using APL and Clipper

Inna N.Luneva
Sergei M.Obraztsov
Alexander L. Shimkevich
Institute of Physics and Power Engineering,
249020, Obninsk
Russia

Abstract

This paper describes the algorithm for the prediction calculation based upon the non-linear algebraic models' self-organization. Special block for a function recognition provides the decrease of the predicted equations exhaustive search at every stage of model building. Another peculiarity is the bootstrap use as supplementary external criterion. The implementation of this algorithm is a complex software system (calculating module is built in APL*PLUS and user-friendly interface is built in Clipper 5.01). This system works as DOS-application for Windows with dynamic data exchange. We present results of testing the system on real currency exchange rate.

Introduction

The problem of response extrapolation on unachievable factor space for observation is very important for scientific investigations, any technological process optimisation, medicine, economics, ecology, and so on. This task has no general solution at present. The results may be obtained with development of computer technology and heuristic methods. The group method of data handling (GMDH) [1] is one of the heuristic methods which are based on the self-organization theory. This theory postulates the existence of the optimal model as a result of model search with advisable criterion use. The traditional regression analysis uses the method of least squares as internal search criterion whereas GMDH is based on the use of an external criterion (the input data is used here, it is not used in parametric model identification).

The heuristic character of models' self-organization is exhibited at the choice of support functions, variables' standardization and so on. So, a method is heuristic if it is based on the use of rules, methods and simplifications that summarize all investigator's experience.

The algorithm of prediction calculation based on models' self-organization is described below.

1. Description of Algorithm for Consecutive Picking Out Trends

1.1. The Main Ideas

It is assumed that all investigated functions are subjected to some deterministic laws and have not structural alteration on the prediction interval. It is also assumed that the data obtained previously reflect all peculiarities of the phenomenon. In this case, the process dynamics can be expressed in the following way:

$$y(t) = \sum_{i=1}^{k} f_i(\Theta, t) \qquad (1)$$

where y is a response value,
t is an independent variable (argument),
f is a prescribed function,
Θ is a parameter vector of model (1),
k is the number of optimal model members.

The optimal model must qualitatively describe the process of "pre-history" in agreement with the least squares criterion (interpolation) and it must provide a satisfactory response on the forecasting interval (extrapolation). The optimal model complication is equal or less than investigated object complication. It is

APL 94 - 9/94 Antwerp Belgium
© 1994 ACM 0-89791-675-1/94/0009..$3.50

impossible to satisfy these conditions simultaneously if all sampling points will be used equally likely, indeed errors will tend to zero unrestrictedly with model complication increase while predicted values will not be stable. To avoid this contradiction, a certain part of sampling points is to be used for the model quality analysis. The algorithm demands to partition the sample for the training and testing. The model is built on the training part of the sample and is verified on the testing one. Partitioning is not a formalized procedure. It is influenced by the experience and intuition of investigator. Moreover, this factor can be decisive in the search for solution.

The set of support models is defined beforehand. The functions involved in (1) are chosen from this set at identification. Then the residual calculation is performed:

$$e_l = Y - \hat{Y}_{l-1}, \quad \hat{Y}_0 \equiv 0 \qquad (2)$$

where $\hat{Y}_l$ is approximation of (1) after l-iteration, $l = 1, \ldots, k$.

This procedure is repeated as long as least squares will become minimal in the testing part of the sample.

1.2. Efficient Algorithm

A set, **U**, of the continuous twice differentiable functions, $\{C^2\}$, is used in the scheme described above as support models:

$$\mathbf{U} = \{f : f \in C^2\} \qquad (3)$$

Obviously, the exhaustive search of all U-functions leads to a significant decrease of computational efficiency. In order to restrict the number of combinations, a special block for recognition of residual type is provided. It is assumed that residuals can be referred to one of five groups of functions:

- increasing convex upwards
- increasing concave
- decreasing convex upwards
- decreasing concave
- oscillating.

The group is selected at random manner in agreement with a 2x2 probability matrix, ***P***. This matrix is obtained as a result of three consecutive residual values step by step analysis. ***P*** is formed in such a manner that:

- each second value is taken from three and it is compared with the preceding value and next one;
- the result is accumulated in the matrix element that is responsible for the concrete function type;
- the matrix is reduced to fulfilling normality condition form

$$\sum_{i,j=1}^{2} p_{i,j} = 1 \qquad (4)$$

The choice probability of certain subset ***u*** from **U** is determined by the corresponding element of probability matrix. Elements of the main diagonal are responsible for increasing and decreasing function types. The degree of oscillation is defined by elements of the secondary diagonal.

The randomization is used in order to keep freedom of choice (this principle is fundamental for the self-organization theory), because the desired function belongs rarely to a unique type. As a rule, it is a mixture of the certain number of types.

All predicted functions are clustered in agreement with this classification, so that the number of computational operations decreases to minimum.

After the class choice, the identification of its composing models follows [2]. An estimation is accomplished successively by three steps:

- approximate definition of the solution existence domain;
- global search procedure (by Monte Carlo method);
- accurate estimation by Marquardt method.

The second factor of the algorithm is the bootstrap [3] used to decrease the sampling bias for the model parameters' estimates. The training (but not testing) part of sampling is multiplied repeatedly and the pseudo-experimental data estimation allows to look at the process from different points of view. It is important especially when an analysis of the non-linear dynamic stochastic processes is carried out and the system analysed is too complicate for supposition about the process statistical characteristics. Analogous problems arise when the number of sampling points is strictly limited and parallel observations are absent. For example, it is impossible to answer a priori whether the rate of exchange data is corrupted by noise and to determine the quantity of this noise . The same can be said about the weather short-term forecasting, etc.

One selects from the bootstrap array the estimate which minimizes the least squares on the testing part of sampling. In this manner, bootstrap comes forward as supplementary external criterion for model quality. This method uses imitative experiments for maximum likelihood keeping (that is to make the estimates obtained agree as closely as possible).

2. Description of Algorithm Implementation

All items described above lead up to the following:

- in connection with the algorithm complication and a large number of calculation operations, the program may be developed only with language providing program maintenance flexibility, with great number of functions with high-level matrix-handling and so on;

- a specialist's experience and intuition must be used for prediction task because they can speed up optimal model search. So developed program should have user-friendly interface. In this case, it will allow to control calculation process, to make input data preparation, to present forecasting results in different manner, to accumulate these results for making prediction precision in the future and so on;

- the possibility should exist to transfer prolonged prediction task to background mode. This will allow to carry out another work at the same time.

Taking into account these considerations, a complex software called "PROGNOSIS" has been developed. It consists of two programs - "Calculation" and "User Interface".

"Calculation" is built in APL*PLUS which makes quick, compact, and elegant realization of the complicated algorithm, provides high computational efficiency and allows to present the process in real-time and graphics mode.

"User Interface" is built in Clipper 5.01 which is specially designed for user-friendly interface development with data base technology use. It provides the following possibilities: hierarchical menu creation; data access, storage and retrieval; data modification and sorting, preliminary data processing and so on.

APL and Clipper are interpreters and this circumstance is a great advantage in our case. Execution of the APL primitive "execute" allows to add new objective functions to the predicting models set. Moreover, it simplifies data exchange between tasks: it is sufficient to pass function numbers and estimates (parameter values) as a prediction result, and the macro operator in Clipper that allows runtime compilation of expressions will transfer this result to the digital form.

Both programs can work as DOS-applications under Windows 3.1 at the same time. Data exchange is transparent for user and is realised by files with an especially developed standard. So, it is look-alike Windows DDE-mechanism. Run-time environment allows background processing of any task. Besides, it allows to install job priority and so on.

3. Program Testing

Complex software "PROGNOSIS" has been employed successfully for the processing of liquid-metal technology experimental data. In particular, the critical points for alkali metals' have been calculated on measurement results of density as function of temperature. In agreement with Fisher's ratio test, these estimates have been verified by independent values of dynamic viscosity. This validation shows that representativeness of the results obtained is higher than the ones from litterature.

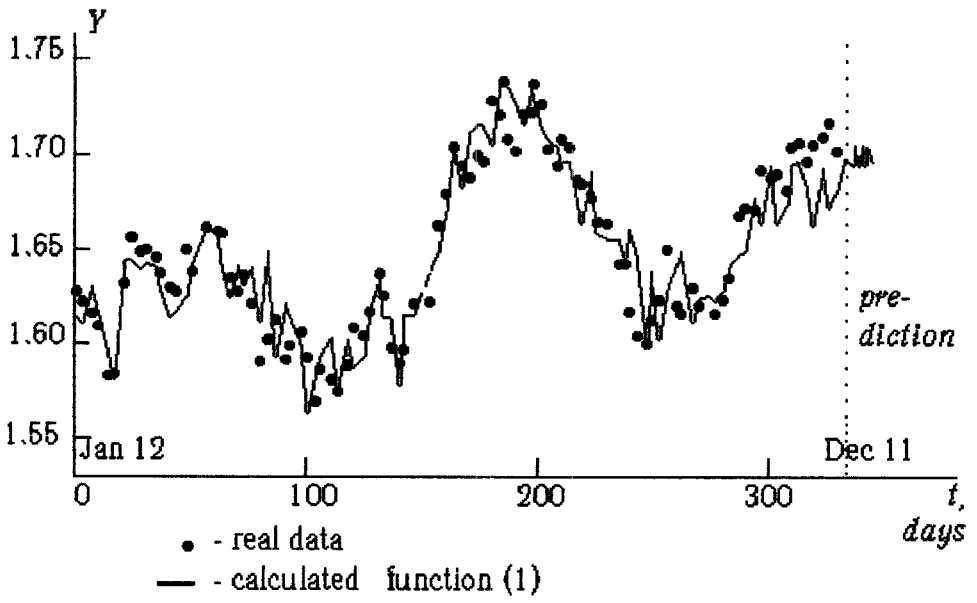

Figure 1. Real data for 1993 and computation results of the optimal model (1).

Here, we show the test of this program on a more interesting example: a prediction of the US dollars versus DM rate of exchange.

In this case, the use of this method is stipulated by great complication of detailed economical simulation. On the other hand, this task solution is very important from the practical point of view.

The model (1) building has been carried out on real Reuters data , Y, in time interval, Δt, from January 13 to December 10, 1993 . It should be noted that sampling is not sufficient for the long-term forecast (more than 20% from observation time), because this data didn't represent low-frequency trends. Nevertheless, it was expected that a short-term forecast (next 2 weeks until December 25) may be built.

The response has been standardized as a preliminary, e.g. the following vector, Y', was analysed instead of initial value Y:

$$Y' = \frac{Y - \bar{Y}}{\sigma_Y} \qquad (5)$$

where Y is the initial rate of exchange,

$\bar{Y}$ is sampling mean,

σ_Y is standard deviation.

Then the sample has been smoothed with a specially developed algorithm [4]. "User Interface" module instantly makes such transformations with two key presses. After that, data has been used as "Calculation" module input, the training part took up 70% of the sample size. As a result of 10 hours computation on IBM PC 386, the model (1) has been obtained and has accumulated 7 members.

It is interesting that the yearly trend is described by the following equation:

$$e_1 = 0.057 \cdot \exp(7.5 \cdot 10^{-3} \cdot t) \qquad (6)$$

Real data and computation results of the optimal model are shown on Fig. 1. Fig. 2 presents the values of relative errors, δ:

$$\delta = \frac{y_r - y_{pr}}{y_{pr}} \cdot 100\% \qquad (7)$$

where y_r, y_{pr} are real and predicted values in the prediction interval respectively .

For the calculation of the short-term forecast precision criterion the following formula is used [1]:

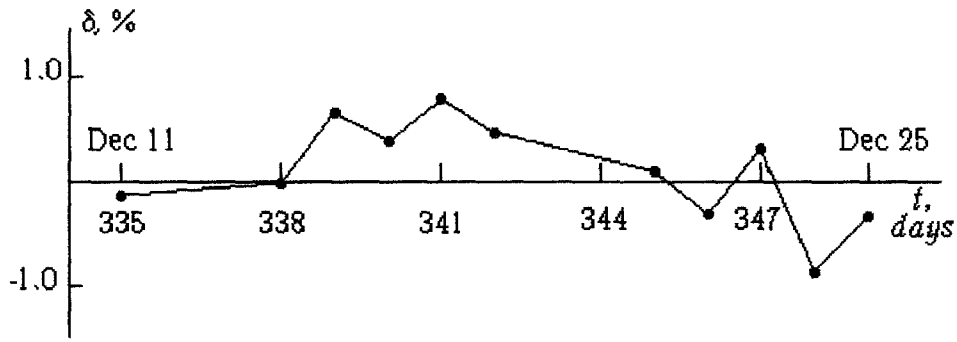

Figure 2. Relative error of the prediction interval.

$$RR = \frac{\sum_{i=1}^{n} (y_{r_i} - y_{pr_i})^2}{\sum_{i=1}^{n} y_{r_i}^2}, \qquad (8)$$

where n is the number of prediction points.

Prediction is good for $RR \leq 0.5$, it is satisfactory for $RR \leq 0.8$, and if $RR > 1$ it is impossible to use the forecasting model. It is $RR < 10^{-2}$ in our case that gives excellent prediction.

The residuals belong to normal distribution according to ω^2-criterion. Sample mean equals to zero according to Student criterion with confidence level equalled to 0.01.

One can state that we have succeeded in identification of determination component of process. Stochastic component can be described by normal distribution N (0, 0.019).

Conclusions

The development of a prediction models' self-organization approach is suggested. It is based on a recognition block and bootstrap application. The program designed on this basis allows to calculate forecasting values for data of various physical nature.

Comfortable user environment is designed on the basis of APL- and Clipper-module integration. The united programmed complex works under Windows control.

Acknowledgements

The authors would like to thank Prof. Victor M. Murogov for management, Prof. Pavel L. Kirillov and Dr. Vyacheslav M. Kuprianov for their critical view on the subject of the paper and useful advices.

References

[1] **J.A.Muller. A.G.Ivachnenko**. Selbstorganisation von Vorhersagemodellen.- Berlin, VEB Verlag Technik, 1984.

[2] **N.R.Draper, H.Smith.** Applied Regression Analysis. Second edition.- New York, John Willey & Sons, 1985.

[3] **B.Efron**. Bootstrap methods: another look at the jackknife. The Annals of Statistics. 1979, Vol.7, N.1.

[4] **V.M.Kuprijanov, O.P.Lucksha, S.M.Obraztcov, A.L.Shimkevich**. Non-linear model parameters' investigations with bootstrap. - Preprint IPPE-2223, Obninsk, Russia, 1991.