Automatic identification of ARIMA models with neural network

Balázs Lénárt

Abstract

The paper investigates an artificial intelligence based demand forecasting method. A neural network driven automatic ARIMA model identification is being introduced. The limitations of the current methods are shown and a new identification concept is presented. It is being discussed that the model identification with a neural network is less sensitive to input errors through its intuitive capability, additionally after a certain number of training steps the algorithm is able to identify time series with unknown characteristics.

Balázs Lénárt

Department of Transport Technology, Faculty of Transportation Engineering, Budapest University of Technology and Economics, Budapest, Hungary, 1111 Budapest, Bertalan Lajos u 2., Hungary
e-mail: lenart@kku.bme.hu

1 Introduction

In the last decade artificial intelligence (AI) methods come into prominence. The main reason is that the artificial intelligent methods are the mathematical models of human thinking and natural laws, therefore, a human-made decision support system (DSS) can behave similar way as an intelligent living being. With this ability the commonly used logistics methods can be developed in different fields such as planning and operation. In some cases the human intelligent can be swapped or complemented with artificial intelligence methods. These methods could be inventory, scheduling, shortest route problems.

The main purpose of the research is to develop a time series analysis method in order to increase the demand forecast accuracy. The examined method is the Autoregressive Integrated Moving Average (ARIMA) model [1], which has outstanding forecast accuracy in case of an auspicious identification. The aim is to show that the automatic ARIMA function identification can be accomplished with an artificial neural network (ANN) and with its learning ability the efficiency of identification is growing.

2 Difficulties of automatic identification

The commonly used identification method is able to detect the proper ARIMA model of theoretical time series [2], but in real conditions we hardly meet such "ideal" series, in most cases the data contain errors. In inventory management it is known that the demand time series contain irregular spikes and gaps (high volume ad hoc orders, stock shortages). These irrelevant values – called outliers – appear (Fig. 1) in the partial autocorrelation (PACF) and autocorrelation (ACF) function making automatic identification impossible.

To eliminate the outliers several methods, such as Bartlett Test [3], Level-1, Level-2 Distance Test [4], Multi Criteria Identification [5], are known, but none of the listed tests is able to identify the ARIMA(p, d, q) model in all cases. Outliers are easily recognized by experts, (in most cases) the human mind is capable to determine the correct p, d and q values. This led to the idea of integrating learning capabilities into the identification algorithm. To achieve this, a neural network may be embedded into
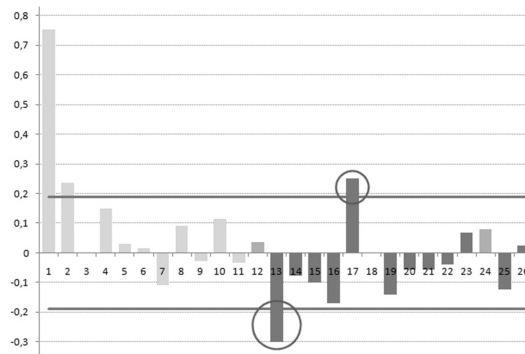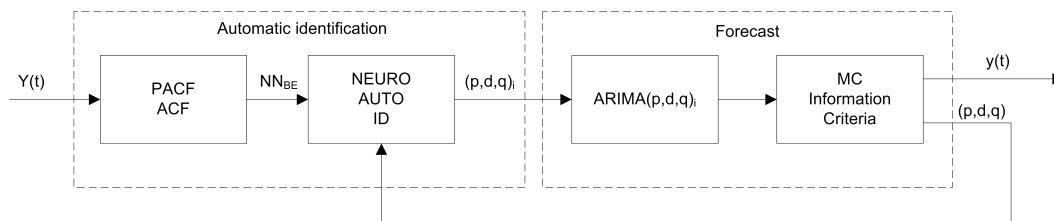
Fig. 1. Outliers in PACF function

Fig. 2. Forecast model

the model identification process.

3 The forecast model structure

The forecasting model can be split into two parts, to an automatic identification module and to a forecasting module (Fig. 2). In the initial step of the first module the PACF and ACF function are created, however, the above mentioned logic identification methods (tests) are changed to learning-capable neural network (NEURO AUTO ID). The outputs of the identification module still remain the parameters of ARIMA(p, d, q) model.

In the early phase of the learning process the ANN provide more parameter sets, but after a short learning period the number of parameter sets may drop dramatically. The initial knowledge is provided by a training dataset, then the (p, d, q) parameters of the best fitting y(t) output is fed back to the neural network.

The second module is responsible for making the forecast. As it was mentioned the identification module in the initial phase provides a large set of (p, d, q) parameters and all of this sets are calculated in this phase. This is called parameter estimation (ϕ, ϑ) which objective is to fit the ARIMA(p, d, q) model to the original time series by least square method. This is a multiparameter optimization problem, therefore, an evolutionary algorithm – e.g. genetic algorithm – is suggested. Out of the results the best fitting parameter set may be chosen. To determine the fitting accuracy several methods are known, these could be simple statistical quantities, such as mean absolute error (MAE), mean absolute percentage error (MAPE), or ARIMA specific criteria, such as final prediction error (FPE) [6], Akaike information criteria (AIC) [7] and normalized Bayesian information criteria (nBIC) [8]. A multi-criteria decision support algorithm – containing the above statistical methods – was developed to select the best fitting model.

4 Parameter estimation with genetic algorithm

H indicates a set which was selected by the identification module. The mathematical structure of the model depends on the (p, d, q) parameters which was determined also by identification module. Using this parameters and the operator equation the analyzed $H(i)$ARIMA model can be resolved. The estimated parameters in an ARIMA(p, d, q) model are the autoregressive (ϕ) and moving average (ϑ) parameters, which calculation is made by a genetic algorithm (GA).

A GA generates solutions to optimization problems using techniques inspired by natural evolution, in this case it finds the best fitting $H(i)$ model by iteratively changed parameter settings. The GA used in the forecast module is a so called floating point genetic algorithm (FPGA). The last step is to save the above calculated values in the actual $H(i)$ set, hence they are getting used in a subsequent phase. This subprocess must be performed on all pre-identified $H(i)$ model. The flowchart of the parameter estimation process is shown schematically in Fig. 3.

5 Integrating neural network into the model

The ANN is a connected mesh of artificial neurons, and it simulates some properties of the biological neural network. The most important property of the ANN is the learning capability. An ANN is capable to change the weights of its internal connections (synapses) while learning, in order to assign the input pattern to the wanted output data. This is particularly useful in applications where the complexity of the data or task makes the design of such a function by hand impractical [9]. The aim of the current research is to show that the automatic identification with a neural network is more efficient than the current methods, it has high error tolerance and with its learning capability it is able to adapt to the constantly changing environment.

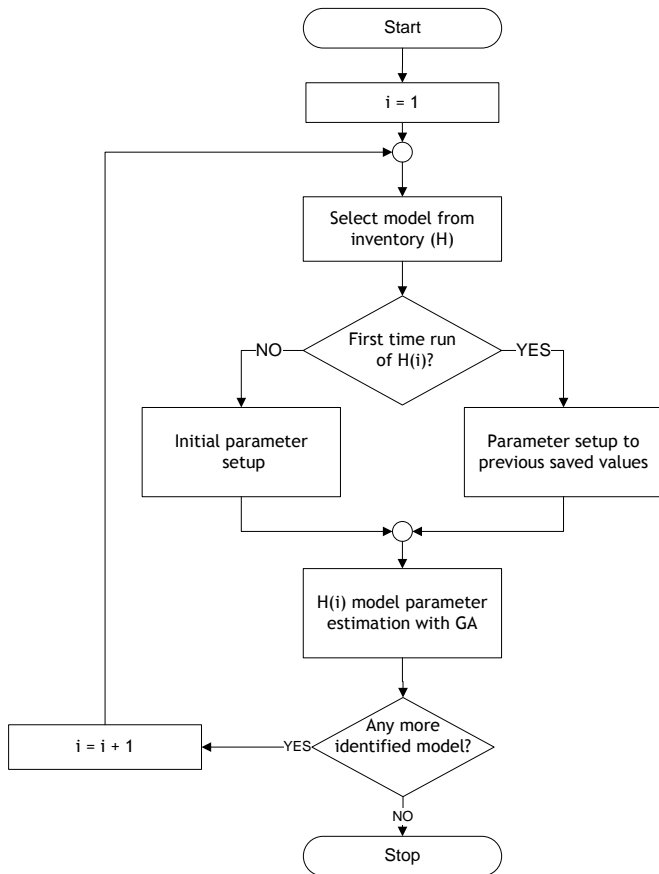During the process of model building it is important to de-

Fig. 3. The parameter estimation process

Fig. 4. Neural Network

termine the inputs, outputs, structure and the learning method of the ANN (Fig. 4). For the current investigation a fully connected ANN was selected. The inputs of the network are the values of the PACF and ACF functions. The number of input neurons depends on the length of the time series and on the trend parameter. If the length of the time series is T and the trend parameter is d then the number of inputs is:

$$NN_{BE} = \sum_{d=0}^{D}(T - d - 1)$$

The number of neurons in the hidden layer is selected by experientially. The output layer contains 6 neurons and the selected models are determined by these neurons in a way that multiple ARIMA(p, d, q) models are being selected.

Then all of the pre-selected models are calculated in the forecast module, after this the best fitting ARIMA(p,d,q) model is being selected by the above mentioned multi-criteria decision making algorithm. The prediction is given by this selected model and it is fed back to the neural network ion the purpose of training. The used training method is the backpropagation algorithm [10].

6 Summary

The automatic ARIMA(p, d, q) model identification described in this article is a result of a continuous research and development. It was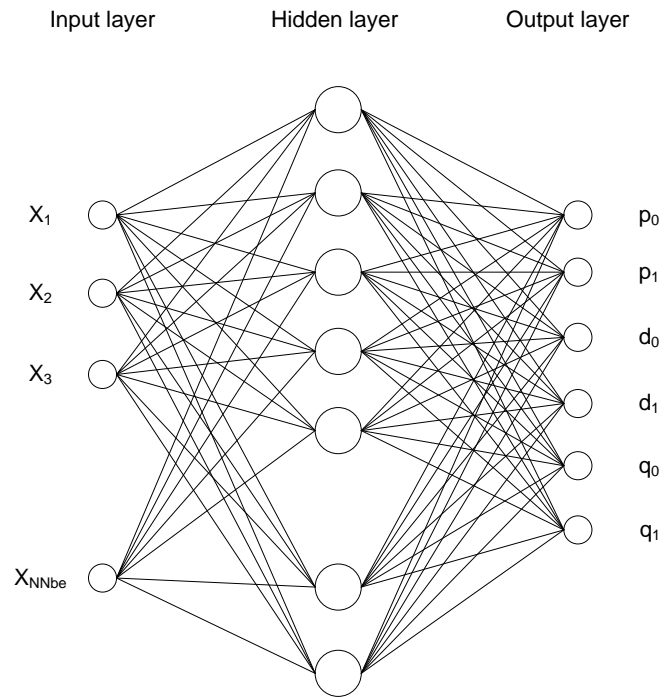 presented that the current identification methods and tests are hardly usable for "non ideal" time series and they are unable to adapt to the constantly changing characteristics of input data. On the other hand the model identification with a neural network is less sensitive to input errors through its intuitive capability, additionally after a certain number of training steps the algorithm is able to identify time series with unknown characteristics. These features make the method capable of integrating it into logistics processes. Of course the presented forecast method is applicable to support other problems. The learning and intuitive capability of the ANN is usable on any fields, where a prompt decision must be done with the support of previously acquired knowledge. It is effectively applicable solving inventory, scheduling, shortest route problems. The accuracy and efficiency of the method highly depends on the previously acquired knowledge, therefore, a great importance must be attached to the planning and operation. If the attention is made then the result is a robust, fast and flexible system wherein the unknown and random events can be treated effectively.

References

1 **Box G, Jenkins G**, *Time series analysis: Forecasting and control*, San Francisco: Holden-Day, 1970.

2 **Péter T, Korcsog A**, *Rakodási rendszerek input-output folyamatainak jellemző periodicitásának vizsgálata*, AUTOMATIZÁLÁS **XIX** (1986), no. 6, 26-33.

3 **Bartlett M S**, *Properties of sufficiency and statistical tests*, Proceedings of the Royal Statistical Society Series A **160** (1937), 268–282.

4 **Tran N, Reed D A**, *Automatic ARIMA Time Series Modeling for Adaptive I/O Prefetching*, IEEE Transactions on parallel and distributed systems **15** (February 2004), no. 2.

5 **Bóna K, Lénárt B**, *Idősor előrejelzés SARIMA módszerrel*, 2009. (Kutatási összefoglaló.

6 **Shibata R**, *Approximate efficiency of a selection procedure for the*

number of regression variables, Biometrika **71** (1984), 43–49, DOI 10.1093/biomet/71.1.43.

7 **Akaike Hirotugu**, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control **19** (1974), no. 6, 716–723.

8 **Liddle A R**, *Information criteria for astrophysical model selection*, available at `http://xxx.adelaide.edu.au/PS{_}cache/astro-ph/pdf/0701/0701113v2.pdf`.

9 **Bishop C M**, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.

10 **Arthur Earl Bryson, Yu-Chi Ho**, *Applied optimal control: optimization, estimation, and control*, Blaisdell Publishing Company or Xerox College Publishing, 1969.