Contents lists available at ScienceDirect

Neurocomputing

A new model selection strategy in time series forecasting with artificial neural networks: IHTS

Serkan Aras *, İpek Deveci Kocakoç

Econometrics Department, Faculty of Economics and Administrative Sciences, Dokuz Eylul University, Izmir, Turkey

A B S T R A C T

Although artificial neural networks have recently gained importance in time series applications, some methodological shortcomings still continue to exist. One of these shortcomings is the selection of the final neural network model to be used to evaluate its performance in test set among many neural networks. The general way to overcome this problem is to divide data sets into training, validation, and test sets and also to select a neural network model that provides the smallest error value in the validation set. However, it is likely that the selected neural network model would be overfitting the validation data. This paper proposes a new model selection strategy (IHTS) for forecasting with neural networks. The proposed selection strategy first determines the numbers of input and hidden units, and then, selects a neural network model from various trials caused by different initial weights by considering validation and training performances of each neural network model. It is observed that the proposed selection strategy improves the performance of the neural networks statistically as compared with the classic model selection method in the simulated and real data sets. Also, it exhibits some robustness against the size of the validation data.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Forecasting time series is a crucial and constantly growing discipline, which includes different areas such as economy, industry, and science. Linear methods like moving averages, exponential smoothing, and ARIMA have dominated the field of time series forecasting for a very long period and found place in real-world applications. Most real-world problems have actually non-linear characteristics, which has drawn attention to non-linear techniques. With the aim of modeling non-linear patterns, some non-linear methods such as Bilinear, Threshold Autoregressive (TAR), and Generalized Autoregressive Conditional Heteroskedastic (GARCH) have been developed. The problem with these model-based techniques is the need to specify a non-linear functional relationship between variables. This property of model-based techniques has limited their applications in some specific problem areas [1].

The need for flexible methods to overcome the necessity of specifying a non-linear relationship before modeling has made some researchers to focus on neural networks because of their remarkable features. Theoretically speaking, it is proved that a neural network with a single hidden layer containing a sufficient number of hidden units can approximate any continuous function

to any degree of accuracy [2,3]. Also, a large number of successful applications of neural networks in time series forecasting show that it is a valuable tool in modeler's toolbox [4].

A large number of comparisons between neural networks and traditional time series forecasting techniques have been made in order to determine the potential power of neural networks in time series forecasting. Even if most of these studies have found that neural networks have superiority over traditional techniques [5–10], there are other studies reporting that traditional techniques outperform neural networks [11–14]. Although it is a known fact that any forecasting technique is not better than all others in all cases [15], the notion that neural networks do not have a certain methodology widely accepted by researchers may lead to these contradictory results. The uncertainty of selection of a particular neural network as a final model for future uses is one of the methodological shortcomings of neural networks.

The classical method divides data sets into training, validation, and test sets, uses validation set to decide when to stop training, and, after the training process, selects the neural network model which gives the smallest error value in the validation set (Early Stopping). Even though the aim of this method is to avoid selecting a neural network model overfitting training data, it is known that data splitting can increase forecast variability [16,17]. When the number of data points is limited, or the effect of the last period on the related series is strong, selection of a large size of validation set might lead to troubles. However, if the size of the

* Corresponding author. Tel.: +90 5552543728.
  E-mail address: serkanaras35@hotmail.com (S. Aras).

validation set is small, the chance of having poor generalization of the neural network, selected by the classic method, might increase. There are other techniques known as regularization to prevent overfitting in the literature [18–21]. The idea behind regularization techniques is to keep the network small by penalizing the network complexity. Hagan et al. [22] demonstrated that the regularization and early stopping methods are, in effect, performing the same operation and reported an approximate equivalence of the two methods. This study proposes a new selection strategy in early stopping method in order to overcome the selection of a neural network that has poor generalization ability caused by the size of the validation data.

The proposed selection strategy (Input, Hidden and Trial Selection – IHTS) is based on the findings obtained by previously conducted studies in which the input units were deemed more important than the hidden units in the performance of neural networks [23–26]. Hence, it first tries to identify the number of input units (past/lagged observations), and then, the number of hidden neurons. After the numbers of input and hidden units are determined, it is decided which trial, among other trials caused by different initial weights, will be selected by taking into account the performances of validation and training sets simultaneously. This step helps to select a neural network good not only at validation data but also at training data. TOPSIS, one of the multi-criteria decision-making techniques, is employed for this selection.

The rest of the paper is organized as follows. The subsequent section discusses the problems related to the determination of a final neural network model. Section 3 mentions the importance of input and hidden units in time series forecasting. The proposed model selection strategy is presented in detail in Section 4. Section 5 provides data sets and experimental design. The obtained results are reported in Section 6. Finally, conclusions and some further discussions are given in Section 7.

2. Model selection problem in neural networks

The selection of the best architecture is important to achieve success in neural networks applications [6,27]. The universal approximation property of neural networks is related to any function that is defined over a certain range. From the forecasting point of view it can be inferred that this property has a limited value. This is because the series under investigation bears some noises at various levels and the points to be forecasted is out of the series that has been used to build the model. Further, a neural network attains the universal approximation property by means of adding a finite number of hidden units to a hidden layer [2,3]. But a high approximation of neural networks to training data with noises means that noises will be learnt or memorized-it is called overfitting by statisticians. The problem regarding neural networks is to violate the usual assumptions about linearity and sampling distribution. This invalidates many attempts intended for establishing statistical tests to select a proper model. Although some steps were taken to deal with this problem, they have failed since distributions do not exist [28].

Some design factors such as selection of input variables, architecture of neural networks, and the size of the training data have high influences on the forecast accuracy of neural networks [4]. Tang et al. [5] stated that neural networks are a promising alternative to traditional techniques for forecasting issues, but they suffer from the problems of specifying optimal topology and parameters. Tang and Fishwick [29] reported that better performances can be reached by using neural networks; however, the lack of systematic procedures and theoretical background in building a neural network model prevent them from happening. The necessity of deciding about many parameters before building a neural network still requires the process of model construction to include a great number of trials and errors. Therefore, numerous neural networks, each of which has different architecture and parameter values, should be built before finding a satisfying model. Deciding on many parameters actually forms the source of flexibility used by neural networks to fit any function. The cost of this flexibility is the time and effort spent by researchers on finding the right combination of parameters.

Lachtermacher and Fuller [23] tried to identify the number of input units by utilizing one aspect of the Box–Jenkins methodology and to specify the number of hidden units with the help of a heuristic rule. The weaknesses of their study are the way of determining input units for a non-linear method by means of linear autocorrelation function and not using a systematic approach for choosing hidden units. Unlike most of the pruning methods for neural networks, Cottrell et al. [30] sought a solution to the problem of architecture selection with statistical techniques based on asymptotic properties of the estimator. Their proposed method depends heavily on some statistical assumptions, and the final model can be affected by the initial model. In addition, they employed Akaike and Bayesian information criteria (AIC and BIC), which were not specially developed for neural networks. Qi and Zhang [31] showed that they were not reliable model selection criteria to choose the model that has the best out-of-sample performance. Similarly, Heravi et al. [32] used BIC to select a final neural network model, and Faraway and Chatfield [33] made use of in-sample error, AIC and BIC in the hope of finding the best combination of input-hidden units.

Anders and Korn [34] proposed five model selection strategies relying on different statistical procedures for neural networks specification. They exploited Wald-test statistics and information criteria, such as the AIC and Network Information Criterion (NIC), in order to eliminate insignificant connections. For selection of hidden units, two tests, one (TLG test) proposed by Terasvirta et al. [35] and the other (White test) by White [36], were used. Both Wald and information criteria tests cannot be theoretically verified in case of irrelevant hidden units. Therefore, it is necessary to conduct a test of hidden units before testing input units. Also, it is possible to use AIC or NIC for the purpose of selecting hidden units. Four strategies were given as: 1) White-Wald; 2) TLG-Wald; 3) AIC-AIC; and 4) NIC-NIC. The fifth strategy was composed of cross validation (CV). They stated that cross validation was the most applicable strategy, among others, in the selection of a neural network model because it does not have any probabilistic assumption and is not affected by identification problems. However, carrying out cross validation for each combination of input and hidden units would be quite costly in terms of computation. Further, it is impossible to test a forecasting model in time series through cross validation. Otherwise, it means that future observations would be used to forecast past observations. The most reasonable way is to utilize the basic cross validation applied in this study.

Egrioglu et al. [37] proposed a model selection criterion that takes some model selection criteria (AIC, BIC) and performance measures (RMSE, MAPE, direction accuracy (DA) and modified direction accuracy [MDA]) into account to find out the best numbers of input and hidden units. They formulated it as weighted information criterion (WIC), all of the components of which are weighted subjectively and linearly. Following this study, Aladag et al. [38] specified the weights of WIC objectively with the help of an optimization approach and named it as AWIC. Both studies did not distinguish between input and hidden units considering their relative importance in a neural network performance and utilized information criteria and performance measures most of which are highly correlated with each other.

Walczak and Cerpa [39] stated that wrong selections of four design factors reported as input variables, learning method, the numbers of hidden layers and units can result in failure or suboptimality. But it is known from many successful applications that a neural network with a single hidden layer is mostly used in