Expert Systems with Applications 41 (2014) 4235–4244

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Neural network ensemble operators for time series forecasting

Nikolaos Kourentzes *, Devon K. Barrow, Sven F. Crone

Lancaster University Management School, Department of Management Science, Lancaster LA1 4YX, UK

ARTICLE INFO

Keywords:
Time series
Forecasting
Ensembles
Combination
Mode estimation
Kernel density estimation
Neural networks
Mean
Median

ABSTRACT

The combination of forecasts resulting from an ensemble of neural networks has been shown to outperform the use of a single "best" network model. This is supported by an extensive body of literature, which shows that combining generally leads to improvements in forecasting accuracy and robustness, and that using the mean operator often outperforms more complex methods of combining forecasts. This paper proposes a mode ensemble operator based on kernel density estimation, which unlike the mean operator is insensitive to outliers and deviations from normality, and unlike the median operator does not require symmetric distributions. The three operators are compared empirically and the proposed mode ensemble operator is found to produce the most accurate forecasts, followed by the median, while the mean has relatively poor performance. The findings suggest that the mode operator should be considered as an alternative to the mean and median operators in forecasting applications. Experiments indicate that mode ensembles are useful in automating neural network models across a large number of time series, overcoming issues of uncertainty associated with data sampling, the stochasticity of neural network training, and the distribution of the forecasts.

1. Introduction

With the continuing increase in computing power and availability of data, there has been a growing interest in the use of artificial Neural Networks (NNs) for forecasting purposes. NNs are typically used as ensembles of several network models to deal with sampling and modelling uncertainties that may otherwise impair their forecasting accuracy and robustness. Ensembles combine forecasts from the different models that comprise them. This paper proposes a new fundamental ensemble operator for neural networks that is based on estimating the mode of the forecast distribution, which has appealing properties compared to established alternatives.

Although the use of ensembles is nowadays accepted as the norm in forecasting with NNs (Crone, Hibon, & Nikolopoulos, 2011), their performance is a function of how the individual forecasts are combined (Stock & Watson, 2004). Improvements in the ensemble combination operators have direct impact on the resulting forecasting accuracy and the decision making that forecasts support. This has implications for multiple forecasting applications where NN ensembles have been used. Some examples include diverse forecasting applications such as: economic modelling and policy making (Inoue & Kilian, 2008; McAdam & McNelis, 2005), financial and commodities trading (Bodyanskiy & Popov, 2006; Chen & Leung, 2004; Versace, Bhatt, Hinds, & Shiffer, 2004; Yu, Wang, & Lai, 2008; Zhang & Berardi, 2001), fast-moving consumer goods (Trapero, Kourentzes, & Fildes, 2012), tourism (Pattie & Snyder, 1996), electricity load (Hippert, Pedreira, & Souza, 2001; Taylor & Buizza, 2002), temperature and weather (Langella, Basile, Bonfante, & Terribile, 2010; Roebber, Butt, Reinke, & Grafenauer, 2007), river flood (Campolo, Andreussi, & Soldati, 1999) and hydrological modelling (Dawson & Wilby, 2001), climate (Fildes & Kourentzes, 2011), and ecology (Araújo & New, 2007) to name a few. Zhang, Patuwo, and Hu (1998) lists multiple other forecasting applications where they have been employed successfully.

NN ensembles are fundamental for producing accurate forecasts for these various applications; hence, improvements in the construction of the ensembles are important. In this paper, the performance of the proposed mode operator is investigated together with the two existing fundamental ensemble operators: the mean and the median. Two different datasets, having in total 3443 real time series, are used to empirically evaluate the different operators. Furthermore, ensembles of both training initialisations and sampling (bagging) are used to investigate the performance of the operators. The proposed operator is found to be superior to established alternatives. Moreover, the robustness and good performance of the median operator is validated. The findings provide useful insights for the application of NNs in large scale forecasting systems, where robustness and accuracy of the forecasts are equally desirable.

The rest of the paper is organised as follows: Section 2 discusses the benefits of NN ensembles and the limitations of the established

* Corresponding author. Address: Department of Management Science, Lancaster University Management School, Lancaster, Lancashire LA1 4YX, UK. Tel.: +44 1524 592911.

E-mail address: n.kourentzes@lancaster.ac.uk (N. Kourentzes).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.12.011

ensemble operators. Section 3 introduces multilayer perceptrons that will be used for this paper and Section 4 discusses the three fundamental ensemble operators and presents the proposed method for mode ensembles. Sections 5 and 6 discuss the experimental design and the results, respectively, followed by a discussion of the findings in Section 7.

2. Forecasting with neural networks

Over the last two decades there has been substantial research in the use of NNs for forecasting problems, with multiple successful applications (Zhang et al., 1998). Adya and Collopy (1998) found that NNs outperformed established statistical benchmarks in 73% of the papers reviewed. NNs are flexible nonlinear data driven models that have attractive properties for forecasting. They have been proven to be universal approximators (Hornik, Stinchcombe, & White, 1989; Hornik, 1991), being able to fit to any underlying data generating process. NNs have been empirically shown to be able to forecast both linear (Zhang, 2001) and nonlinear (Zhang, Patuwo, & Hu, 2001) time series of different forms. Their attractive properties have led to the rise of several types of NNs and applications in the literature (for examples, see Connor, Martin, & Atlas, 1994; Efendigil, Önüt, & Kahraman, 2009; Khashei & Bijari, 2010; Zhang et al., 1998).

While NNs powerful approximation capabilities and self-adaptive data driven modelling approach allow them great flexibility in modelling time series data, it also complicates substantially model specification and the estimation of their parameters. Direct optimisation through conventional minimisation of error is not possible under the multilayer architecture of NNs and the back-propagation learning algorithm has been proposed to solve this problem (Rumelhart, Hinton, & Williams, 1986), later discussed in the context of time series by Werbos (1990). Several complex training (optimisation) algorithms have appeared in the literature, which may nevertheless be stuck in local optima (Hagan, Demuth, & Beale, 1996; Haykin, 2009). To alleviate this problem, training of the networks may be initialised several times and the best network model selected according to some fitting criteria. However, this may still lead to suboptimal selection of parameters depending on the fitting criterion, resulting in loss of predictive power in the out-of-sample set (Hansen & Salamon, 1990). Another challenge in the parameter estimation of NNs is due to the uncertainty associated with the training sample. Breiman (1996a) in his work on instability and stabilization in model selection showed that subset selection methods in regression, including artificial neural networks, are unstable methods. Given a data set and a collection of models, a method is defined as unstable if a small change in the data results in large changes in the set of models.

These issues pose a series of challenges in selecting the most appropriate model for practical applications and currently no universal guidelines exist on how best to do this. In dealing with the first, the NN literature has strongly argued, with supporting empirical evidence, that instead of selecting a single NN that may be susceptible to poor initial values (or model setup), it is preferable to consider a combination of different NN models (Barrow, Crone, & Kourentzes, 2010; Ben Taieb, Bontempi, Atiya, & Sorjamaa, 2012; Crone et al., 2011; Hansen & Salamon, 1990; Versace et al., 2004; Zhang & Berardi, 2001). Naftaly, Intrator, and Horn (1997) showed that ensembles across NN training initialisations of the same model can improve accuracy while removing the need for identifying and choosing the best training initialisation. This has been verified numerous times in the literature (for example, see Zhang & Berardi, 2001). These ensembles aim at reducing the parameter uncertainty due to the stochasticity of the training of the networks. Instead of relying on a single network that may be stuck to a local minima during its training, with poor forecasting performance, a combination of several networks is used. In the case of uncertainty about the training data, Breiman (1996b) proposed Bagging (Bootstrap aggregation and combination) for generating ensembles. The basic idea behind bagging is to train a model on permutations of the original sample and then combine the resulting models. The resulting ensemble is robust to small changes in the sample, alleviating this type of uncertainty. Recent research has lead to a series of studies involving the application of the Bagging algorithm for forecasting purposes with positive results in many application areas (Chen & Ren, 2009; Hillebrand & Medeiros, 2010; Inoue & Kilian, 2008; Langella et al., 2010; Lee & Yang, 2006). Apart from improving accuracy, using ensembles also avoids the problem of identifying and choosing the best trained network.

In either case, neural network ensembles created from multiple initialisations or from the application of the Bagging algorithm, require the use of an ensemble combination operator. The forecast combination literature provides insights on how to best do this. Bates and Granger (1969) were amongst the first to show significant gains in forecasting accuracy through model combination. Newbold and Granger (1974) showed that a linear combination of univariate forecasts often outperformed individual models, while Ming Shi, Da Xu, and Liu (1999) provided similar evidence for nonlinear combinations. Makridakis and Winkler (1983) using simple averages concluded that the forecasting accuracy of the combined forecast improved, while the variability of accuracy amongst different combinations decreased as the number of methods in the average increased. The well known M competitions provided support to these results; model combination through averages improves accuracy (Makridakis et al., 1982; Makridakis & Hibon, 2000). Elliott and Timmermann (2004) showed that the good performance of equally weighted model averages is connected to the mean squared error loss function, and under varying conditions optimally weighted averages can lead to better accuracy. Agnew (1985) found good accuracy of the median as an operator to combine forecasts. Stock and Watson (2004) considered simple averages, medians and trimmed averages of forecast, finding the average to be the most accurate, although one would expect the more robust median or trimmed mean to perform better. On the other hand, McNees (1992) found no significant differences between the performance of the mean and the median. Kourentzes, Petropoulos, and Trapero (2014) showed that combining models fitted on data sampled at different frequencies can achieve better forecasting accuracy at all short, medium and long term forecast horizons, and found small differences in using either the mean or the median.

There is a growing consensus that model combination has advantages over selecting a single model not only in terms of accuracy and error variability, but also simplifying model building and selection, and therefore the forecasting process as a whole. Nonetheless, the question of how to best combine different models has not been resolved. In the literature there are many different ensemble methods, often based on the fundamental operators of mean and median, in an unweighted or weighted fashion. Barrow et al. (2010) argued that the distribution of the forecasts involved in the calculation of the ensemble prediction may include outliers that may harm the performance of mean-based ensemble forecasts. Therefore, they proposed removing such elements from the ensemble, demonstrating improved performance. Jose and Winkler (2008) using a similar argument advocated the use of trimmed and winsorised means. On the other hand, median based ensembles, are more robust to outliers and such special treatment may be unnecessary. However, the median, as a measure of central tendency is not robust to deviations from symmetric distributions. The median will merely calculate the middle value that separates the higher half from the lower half of the dataset, which is not