International Journal of Forecasting 32 (2016) 1120–1137

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Cross-validation aggregation for combining autoregressive neural network forecasts

Devon K. Barrow[a,*], Sven F. Crone[b]

[a] School of Strategy and Leadership, Coventry University, Priory Street, Coventry CV1 5FB, United Kingdom

[b] Lancaster University Management School, Department of Management Science, Lancaster LA1 4YX, United Kingdom

ARTICLE INFO

Keywords:
Forecast combination
Bootstrapping
Monte Carlo
Time series
Cross-validation

ABSTRACT

This paper evaluates *k*-fold and Monte Carlo cross-validation and aggregation (crogging) for combining neural network autoregressive forecasts. We introduce Monte Carlo crogging which combines bootstrapping and cross-validation (CV) in a single approach through repeated random splitting of the original time series into mutually exclusive datasets for training. As the training/validation split is independent of the number of folds, the algorithm offers more flexibility in the size, and number of training samples compared to *k*-fold cross-validation. The study also provides for crogging and bagging: (1) the first systematic evaluation across time series length and combination size, (2) a bias and variance decomposition of the forecast errors to understand improvement gains, and (3) a comparison to established benchmarks of model averaging and selection. Crogging can easily be extended to other autoregressive models. Results on real and simulated series demonstrate significant improvements in forecasting accuracy especially for short time series and long forecast horizons.

0. Introduction

Improving the accuracy of univariate time series forecasts remains important in many disciplines, from environmental sciences to business and finance. The approach of combining multiple forecasts has shown particular promise (Clemen & Winkler, 1986; Timmermann, 2006), as has been shown by various empirical studies (Aksu & Gunter, 1992; Clements & Hendry, 2007; Jose & Winkler, 2008; Kourentzes, Barrow, & Crone, 2014; Macdonald & Marsh, 1994; Stock & Watson, 2004) and objective forecasting competitions (Makridakis et al., 1982; Makridakis & Hibon, 2000). The traditional approaches to forecast combination typically involve a set of independent, pre-specified forecasts from different algorithms, which are combined in a second step using a variety of different weighting schemes.

As an alternative to combining the predictions of different algorithms, research in machine learning for predictive classification routinely applies repeated subsampling of the dataset on which a single algorithm is parameterised, creating diversity in the data rather than in the algorithms. The most widely studied methods, bagging (Breiman, 1996a) and *k*-fold cross-validation ensembles (Krogh & Vedelsby, 1995), use different data resampling techniques, namely bootstrapping and cross-validation (CV), respectively, to actively create diverse estimates of the same base learner algorithm for successive combinations of the predictions. Their success in improving the performance and robustness of predictions for classification has been proven empirically in a large number of research studies (see e.g. Dietterich, 2000; Zhou, Wu, & Tang, 2002), with their wide use being reflected in published textbooks

* Corresponding author.
E-mail address: devon.barrow@coventry.ac.uk (D.K. Barrow).

http://dx.doi.org/10.1016/j.ijforecast.2015.12.011
0169-2070/

(see e.g. Perrone & Cooper, 1992) and their availability in standard software packages (see e.g. Matlab and Salford Predictive Modeler Software Suite).

Despite the fact that both methods have been extended to regression in general, and time series forecasting in particular, this class of algorithms has received relatively little attention in forecasting research. While bagging has been assessed in select studies, it was not until recently that cross-validation for time series forecast combination was studied by Donate, Cortez, Sanchez, and de Miguel (2013) and Sorić and Lolić (2013), with promising results. However, both studies were constrained to the variant of k-fold cross-validation, applying a fixed and predetermined number of subsamples in order to create diversity. In contrast, the use of Monte Carlo cross-validation, which combines the benefits of both cross-validation and bootstrapping – repeated random sampling with replacement – in a single approach, has largely been ignored for forecast combination.

In this study, we use cross-validation for combining autoregressive forecasts. The forecast combination averages over a set of forecast models that are trained using mutually exclusive cross-validation replicates sampled from a given learning set. Within the general framework of cross-validation and aggregating, or crogging for short, we introduce a new method of forecast combination, Monte Carlo crogging, and compare it with k-fold crogging and bagging. Thus, the contributions of this research study are fourfold: (1) the first application of Monte Carlo cross-validation to forecast combination; (2) the first systematic empirical evaluation of different cross-validation approaches and bagging across data conditions of time series length and equal number of samples, using a simulated study on both linear and nonlinear data, as well as empirical data; (3) an assessment of performances in terms of a bias and variance decomposition of the mean squared error (MSE) of the forecasts; and (4) a comparison of cross-validation with bagging and established benchmark methods of model averaging and model selection, utilising the 111 time series of the NN3 competition (Crone, Hibon, & Nikolopoulos, 2011).

This paper is organised as follows. Section 2, reviews the literature on forecast combination, error estimation and data sampling, linking the three main areas of this research. In Section 3, we describe the application of bootstrapping and cross-validation for forecast combination through bagging and the proposed crogging framework. We describe several crogging strategies, including the proposed combination based on Monte Carlo cross-validation, and provide some theoretical insights into crogging to explain why it should be an effective strategy for forecast combination. In Section 4, we use an extensive simulation to evaluate the differences between crogging and bagging in terms of bias and variance, and for varying combination sizes and time series lengths, while Section 5 presents the results of an empirical evaluation based on data from the NN3 competition. The final section provides a summary and concluding comments.

1. Forecast combination, error estimation and data sampling

In the 50 years since the seminal paper on forecast combination by Bates and Granger (1969), the majority of the papers in the field have resorted to combining the results of multiple forecast models specified previously, or multiple training initializations thereof, where each one is parameterized on the same complete set of learning data. In contrast, recent methods based on bootstrapping and cross-validation have focused on model estimation and the active creation of diverse predictions over which to average. In this research, we focus on cross-validation that was originally developed for estimating prediction errors and the facilitation of model selection. While our interest is in forecast combination, most current research on cross-validation is in the model selection literature (see the review by Arlot & Celisse, 2010). Here, the estimation of predictive accuracy is important, both for evaluating the accuracy of statistical models and for selecting the final model.

The statistical resampling technique of cross-validation (CV) assesses how well the results of a statistical estimate will generalize to an independent data set (Stone, 1974). The out-of-sample predictive accuracy is estimated by splitting the original data repeatedly into a training set for estimating the model, and a validation set for estimating the error in the predictions. This has the attractive feature that it produces nearly unbiased estimates of the prediction error, and provides a more representative estimation of the true ex ante performance of the model (Efron, 1983; Kohavi, 1995). The technique is used most popularly in out-of-sample evaluations with a single hold-out dataset (Tashman, 2000) and in specific application areas, such as climate forecasting (Michaelsen, 1987) and financial forecasting with statistics and neural networks (Clarida, Sarno, Taylor, & Valente, 2003; Hu, Zhang, Jiang, & Patuwo, 1999; Wolff, 1987). Despite the advantages of the approach, several research studies have also pointed out its limitations. For example, the advantage of obtaining unbiased estimates is known to fail when the number of models grows exponentially with the number of observations. Birgé and Massart (2007) and Hardle and Marron (1985) showed that cross-validation was prone to failure in the presence of outliers. Hart and Wehrly (1986) proved that cross-validation leads to overfitting for positively correlated data (see also Altman, 1990; Hart, 1991; Opsomer, Wang, & Yang, 2001), although Burman and Nolan (1992) later showed it to be asymptotically optimal for stationary Markov process, though only within a specific framework. Less than persuasive early results were also obtained in the case of leave-one-out cross-validation, albeit for error estimation rather than forecast combination (see also the results of Burman, Chow, & Nolan, 1994).

Recent research on cross-validation for time series forecast combination, while very limited, has produced promising results. Recently, Donate et al. (2013) used a weighted k-fold cross-validation scheme to generate neural network ensembles when predicting six real-world time series, and found an improvement in accuracy for short and medium-length series relative to Holt–Winters' exponential smoothing. At around the same time, Sorić and Lolić (2013) proposed the use of the leave-h-out cross-validation (jackknife) combination for the time series forecasting of euro area (EA) inflation, following Hansen and Racine's (2012) work on jackknifing and model averaging. While the approach did not seem to offer any