Neural Network Ensemble: Evaluation of Aggregation Algorithms in Electricity Demand Forecasting

Saima Hassan, Abbas Khosravi and Jafreezal Jaafar

Abstract—This paper examines and analyzes different aggregation algorithms to improve accuracy of forecasts obtained using neural network (NN) ensembles. These algorithms include equal-weights combination of Best NN models, combination of trimmed forecasts, and Bayesian Model Averaging (BMA). The predictive performance of these algorithms are evaluated using Australian electricity demand data. The output of the aggregation algorithms of NN ensembles are compared with a Naive approach. Mean absolute percentage error is applied as the performance index for assessing the quality of aggregated forecasts. Through comprehensive simulations, it is found that the aggregation algorithms can significantly improve the forecasting accuracies. The BMA algorithm also demonstrates the best performance amongst aggregation algorithms investigated in this study.

I. INTRODUCTION

A NN ensemble is a combination of multiple predictive models, used as a single model to achieve better predictive performance. The ensemble forecasting, is thus a combined function of all member models. It is a more powerful way to improve model accuracy than the single best model. Considering multiple models and aggregating their forecast may also reduce the risk of selecting a single bad model. Due to the promising ability of neural network (NN) in learning nonlinear and hidden patterns in data, it has been widely used in forecasting, classification, and data analysis [1], [2], [3]. Despite this popularity, NN generalization power is not satisfactory in many cases [4]. The generalization in the NN can be harmed by a number of factors including parameter initialization, inappropriate network topology, and improper setting of training process parameters. The generalization power of a set of individual NNs can be greatly improved using ensembles [5]. Since the pioneering work on forecast combinations [6], the superiority of the combined forecasts over individual forecasts have been demonstrated by many researchers using different techniques and datasets [7], [8], [9]. NN field has also benefitted from the forecast combination methods. Ensemble modeling [7] and thick modeling [8] are two parallel combining approaches and have already been applied to NN ensemble to improve forecasting performance. Combining the forecasts generated by NNs is often performed by majority (classification) or by simple averaging (in regression) of the outputs coming from individual networks [1]. A weighted ensemble of the networks is utilized which handles the accuracy and/or contribution of each network differently [10], [11]. Weighting has the added advantage of placing an increased weight on NN models with statistically better performance. This in turn allows to increase the combined forecast’s prediction power.

Roli et al. [12] divided the forecasts combination scheme into two types of techniques, fixed and trained. Fumera et al. [13] introduced a theoretical and experimental analysis of linear combiners for classifier combining. That the linear combiners’ performance is dependent on the individual classifiers’ performance and on the correlation between their outputs is shown in the theoretical analysis. More specifically, the authors believed that the enhancement is achieved by utilizing a weighted average rather than a simple average combining rule. Authors in [14] presented the comparisons of the different ensemble combiners previously trained with Adaboost and Aveboost. The comparisons were performed to see if the most suitable way to join boosting ensembles is the boosting combiner. In [15], a voting-averaged method was introduced to create an ensemble combination to handle tasks involving regression. The performance of heterogeneous ensemble with multi-factors approach is evaluated with either the mean or median combination [16]. It has been noted that the ensemble performance based on a single network shows continued stabilization over a variety of ensemble sizes. While the median is the preferred combination method for heterogeneous networks ensemble across noise levels and ensemble sizes, but the mean is more volatile. The scheme of combining forecasts using short-term and long-term forecasting is investigated in [17].

Electricity demand is one of the key parameter required for estimating the amount of additional capacity required to ensure a sufficient supply of energy. Electricity demand forecasts can be used to control the generation and distribution of electrical power more efficiently. Therefore, to better plan the electric energy production, the importance of knowing the electricity demand in advance is obvious and hence the development of an effective and timely forecasting model. To incorporate the nonlinearity of electricity demand, NNs have received substantial attentions in electricity demand forecasting with good performance reported [18], [19], [20]. It has been applied for short-term forecasting in the last two decades with varying degrees of success. Inspired by the working principle of the NN models and the precision obtained through forecasts combination approaches, various
aggregation algorithms of NN ensembles are applied here for electricity demand forecasting.

The rest of this paper is organized as follows: Section II describes the proposed methodology of designing NN ensembles and their aggregation algorithms. The simulation results of these algorithms for NN ensembles are demonstrated in Section III. Finally, Section IV concludes the paper with some remarks and guidelines for future work.

II. FORMING ENSEMBLES

The foundation for the basis of the ensemble technique is finding methods to exploit the data found in redundant NN models rather than just ignoring the information. The procedure for forming ensembles includes two main phases. Firstly, several individual NN models are created and developed using the training samples. These networks form the base of an ensemble. Secondly, forecasts generated using individual NN models are mixed by applying different aggregation algorithms. The Fig. 1 shows the proper sequence in the process of forming ensembles.

A. Splitting Data

The set of samples are divided into three subsets: training ($D_{train}$), validation ($D_{vald}$), and test ($D_{test}$) sets. $D_{train}$ and $D_{vald}$ are used in the training stage of NN models. The $D_{test}$ is used to evaluate the predictive performance of the aggregation algorithms.

B. Construction of Individual Neural Network Models

A number of diversification strategies have been adopted by researchers to enhance the complementarities of the individual members of the neural networks ensemble. Sharkey [21] suggested neural networks categorization that could be created to show diversity with influence put on one of four elements; these four elements are the initial weights, the training data, the training algorithm and the network architecture that is utilized. In this study NN models with different structures are randomly initialized to maximize the diversification of models used for forecasting. Different structures are obtained by changing the number of neurons in the hidden layers. NN models are then trained using $D_{train}$. The Levenberg-Marquardt algorithm is applied for training to guarantee a fast learning speed. To force the NN to have smaller weights and biases, the MSE with regularization $mse_{reg}$ is used. The number of neurons for multiple NN models with five different structures are obtained as:

$$h = c_1*\epsilon(1, n_{total}/5), c_2*\epsilon(1, n_{total}/5) \ldots c_5*\epsilon(1, n_{total}/5)$$

(1)

where $c_1, c_2, \ldots c_5$ represent the classes of neurons and $\epsilon$ indicates a vector of ones. The five structures are taken as of sizes 5, 7, 9, 15, and 17 neurons in the hidden layer, respectively. Assuming $n_{total} = 100$, the five different structures of NN models mean that each class consists of 20 NN models.

C. Forecast Combination

Three algorithms are considered here for combining forecasts.

1) Equal-weights combination of Best NN models (BEST-comb): Arithmetic averaging is the simplest method for forecast combination. This method performs well empirically. It assigns equal weights to individual forecasts (thus termed equal-weights forecasts combination) and often generates reliable forecasts [22]. The method is based on taking the arithmetic mean of individual forecasts for period $t$. Its mathematical representation is given below:

$$\hat{y}_{c,t} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{i,t}$$

(2)

where $\hat{y}_{c,t}$ is the combined forecast, $\hat{y}_{i,t}$ is the individual forecast obtained from the $i$th model.

Consider that $n_{total}$ and $n_{best}$ indicate the number of NN models used in analysis and the number of best performing NN models, respectively. To select $n_{best}$ forecasts among the $n_{total}$ NN models, forecasting is performed with these $n_{total}$ NN models for samples in $D_{vald}$ [23]. Performance of these models is examined through calculation of MAPE and are recorded as $MAPE_{BEST_{vald}}$. The $n_{total}$ NN models are ranked based on their $MAPE_{BEST_{vald}}$ values and the first $n_{best}$ NN models are selected as the superior NN models in term of the generalization power. These NN models are indicated by $NN_{best,i}$ where $i = 1, \ldots, n_{best}$. Other remaining NN models are then discarded.

The $NN_{best,i}$, $i = 1, \ldots, n_{best}$ are then used to forecast samples in $D_{test}$. These forecasts are referred to as the best forecasts. It is important to note that these samples have not been used for training and selection (validation) of NN models. MAPE is also calculated for these forecasts and is recorded as $MAPE_{BEST_{test}}$ for later comparison. The

Fig. 1. The procedure for forming an ensemble of individual NN models.

![Diagram](image-url)
best forecasts generated using \( NN_{\text{best},i} \) are then combined through a simple averaging method. Then MAPE is calculated for this new set of combined forecasts and is called \( MAPE_{\text{BESTcomb}} \).

The proposed method for generating superior forecasts is straightforward. It combines results generated by filtered NN models to reach better forecasts. However, its simplicity does not mean it is not powerful in term of the forecasting accuracies. In fact, the proposed method compensates weakness of individual NN models by aggregating their forecasts and generating a combined one. It has also been investigated that the performance of the simple average largely depends on the relative size of the variance of the forecast errors associated with different forecasting models [24]. This will be appropriate when these models have the same forecast error variance [25].

2) Combination of Trimmed Forecasts (TRIMcomb): Individual forecasts may have large errors because of miscalculations, errors in observations, or misunderstandings. The presence of any outlying forecasts would have undue influence on the combined forecasts if they are averaged. Authors in [26] recommend using trimmed mean by discarding the extreme forecasts. Granger et al. [8] describe the trimming method as “it consists of ranking the predictions \( \hat{y}_{t+h|i}^{(j)} \) using model \( j \) from smallest to largest and removing the 100\( \alpha \)% largest and smallest, leaving 100(1 − 2\( \alpha \)) remaining to be averaged.”

The method of trimmed mean was first proposed in [27]. It consists of three steps: ordering the sample, trimming the tails of a histogram, and averaging the rest. Assume that \( \alpha \) indicates the amount of trimming and that \( \alpha \in [0, 100] \). To determine the optimal amount of trimming for the \( n_{\text{total}} \) NN models, forecasts are obtained using samples in \( D_{\text{valid}} \). With the purpose of calculating an \( \alpha \)-trimmed mean, forecasts in each row of \( Y_{\text{valid}} \) are first ordered from small to large. Then, the forecast trimming number is determined as per the amount of trimming: \( n_{\text{trim}} = \frac{\alpha}{100} n_{\text{total}} \). \( I_{\alpha} \) is defined as \( I_{\alpha} = \{ \frac{n_{\text{total}} + 1}{2}, \frac{n_{\text{total}} + 2}{2}, \ldots, n_{\text{total}} - \frac{n_{\text{total}}}{2} \} \). Finally the \( \alpha \)-trimmed mean is calculated with the Eq. 3 for all forecasts in the \( i^{th} \) row of \( Y_{\text{valid}} \):

\[
\hat{y}_{i}^{\alpha} = \frac{1}{n_{\text{total}} - n_{\text{trim}}} \sum_{j \in I_{\alpha}} \hat{y}_{i,j}
\]

where \( i = 1, \ldots, n_{\text{valid}} \). Once \( \alpha \)-trimmed means are calculated for all samples in \( D_{\text{valid}} \), the MAPE is recorded as \( MAPE_{\text{TRIMvalid}} \) using Eq. 4.

\[
MAPE^{\alpha} = \frac{1}{n_{\text{valid}}} \sum_{i=1}^{n_{\text{valid}}} \left| \hat{y}_{i}^{\alpha} - y_{i} \right|
\]

where \( y_{i} \) is the \( i^{th} \) target value.

The procedure described above is repeated for \( \alpha \) values in \( [\alpha_{\text{min}}, \alpha_{\text{max}}] \). The optimal value of \( \alpha \), called \( \alpha_{\text{opt}} \), is the one that leads to the minimum value of \( MAPE^{\alpha} \). Once a decision about the best trimming amount (\( \alpha_{\text{opt}} \)) is made, samples in \( D_{\text{test}} \) are forecasted and trimmed appropriately using (4). \( MAPE_{\text{TRIMcomb}} \) is calculated for combined forecasts.

Trimmed mean forecast resolves the dispute against simple average where the extreme scores drag away mean from the center. Theoretical and simulation studies indicate that the trimmed mean is a better pointer of the center of the distribution, as the distribution is almost more symmetric than the non-trimmed distribution. Hence little forecasts will be lost while using trimmed means and result in significantly good forecasts when sampling from a thick or heavy tailed distribution [28], [29]. The major disadvantage of the trimmed mean is that it guards against outliers at the expense of ignoring lots of data. Another disadvantage is that the use of trimmed means is not as desirable for inference as conventional means [30].

3) Combination through the Bayesian Model Averaging (BMAcomb): The Bayesian model averaging presents a systematic method of combining predictive distributions from a diverse group of models. The BMA is helpful in analyzing the uncertainty and scrutinizing the robustness of two or more competing models [31]. The BMA generates an averaged model, weights the individual forecasts based on their posterior model probabilities, and assigns higher weights to the models that best fit the data. A Bayesian combination has been applied to the forecasts from three NN-based models to wind speed forecasting [32]. The method is also applied to combine the forecasts obtained from a number of NN models trained using genetic algorithm [33].

Let \( n_{\text{total}} \) and \( b \) represents the number of NN models used in analysis and the estimated coefficients/weight for the BMA model respectively. This paper follows the method described in [34] for calculating the coefficients/weights for the combination method. The structure exercised for the standard BMA estimation is generalized in the BMA methodology developed in [34] by introducing the distinction between focus and auxiliary regressors. The explanatory variables present in the model are called focus regressors and other additional explanatory variables that are of less certain are termed as auxiliary regressors.

Like others Bayesian estimators, this estimator combines prior beliefs on the unknown elements of the model with the extra information coming from the data. The simple likelihood function, the prior distributions on the regression parameters of model \( M_{j} \), and the prior distributions on the model space are the key elements of this method. Since we are not considering any auxiliary variables, hence \( k_{2} = 0 \) (where \( k_{1} \) and \( k_{2} \) indicates the number of focus and auxiliary regressors). Let \( M = \{ M_{1}, \ldots, M_{j} \} \) indicates the model space of \( n_{\text{total}} \) NN models used for forecasting \( Y \) with \( D_{\text{valid}} \) and \( f_{j} \) the forecasts of model \( j \). According to the law of total probability, the probability density function of the BMA probabilistic forecast \( y \) is an average of the posterior distributions \( p(y | M_{j}, D_{\text{valid}}) \) given by individual models \( M_{j} \), weighted by their posterior probabilities \( w_{j} = p(M_{j} | D) \) [32], [35].
Following the assumption made by Magnus [34] and proposed by Zellner and Fernandez [36] and [37] the prior variance $V_j$ excluding auxiliary variable can be given as:

$$V_j^{-1} = g_j M 1$$

where $g = 1/max(n, k_2^2)$ is a constant scalar for each model $M_j$. Since it is assumed that $k_2 = 0$, then no model selection takes place [34] and $M 1 = X1(X1^T X1)^{-1}X1^T$ where $X1$ is $n \times k_1$ and $n$ is the length of validation data outputs. After all these estimation it will generate a vector of coefficient estimates $b$, and standard errors associated with $b$.

The forecasts for all individual $n_{total}$ NN models for $D_{test}$ are generated with the weights obtained in the last section using Eq. 7 and the method is termed as $BMAcomb$:

$$\hat{y}_{BMA}i,j = \hat{y}_{D_{test}} \times b$$

Then MAPE is calculated for the combined forecasts and is named $MAPE_{BMAcomb}$.

III. CASE STUDIES AND RESULTS

A. Data and Experiment Procedure

Monthly demand data from Australian Energy Market Operator (AEMO) for Victoria region is used as a case study. Twelve data sets of the electricity demand data for year 2011 is considered which are recorded after every 30 minutes interval. A potentially critical issue for the performance of the NN is the selection of input variables to the network. The partial autocorrelation function (PACF) is applied on electricity demand data to select the set of inputs to the NN models. Lagged values ($nLags$) of the twelve months are determined through trial and error. Fig. 2 displays the plot of partial autocorrelation for the electricity demand data-set at the 95% confidence level.

The Matlab `premmx` function is applied here to normalize the inputs and target so that they fall in the interval [-1,1]. Normalization ensures that the distribution of values for each net input and output is roughly uniform. Each data set is then divided into three subsets, 60% for training of NN models, 20% for validation, and 20% for testing the performance of the proposed method. 100 single layer NN models are trained using the samples in $D_{train}$ ($n_{total} = 100$).

$\alpha$ in the $TRIMcomb$ method indicates the amount of trimming. The best trimming amount, $\alpha_{opt}$, is determined in $TRIMcomb$ method in the interval [5%,95%] with a 5% increment. The samples in $D_{test}$ are the forecasts with this trimmed amount. The $BMAcomb$ follows the method implemented by Mangus et al. [34] for calculating the coefficients/weights for combination method. The data-file mentioned is created as a matrix containing samples of the validation data-set in the first column ($n \times 1$) followed by the forecasts generated through $n_{total}$ NN models ($n \times k_1$) with $D_{vald}$. Where $k_1$ is the number of $n_{total}$ NN models in this paper.

A $Naive$ approach is used as a benchmark for aggregation algorithms of NN ensembles. The $Naive$ approach states that the demand forecast for a particular time/day equals to the same period’s previous value. In this paper, 24h ahead forecasting of the electricity demand is focused. Thus a variable time-unit-ahead (TUAhead) is defined as 48 for this study. Table I summarizes various parameters used in experiments conducted in this study.

B. Performance Metric

The forecasting performance is assessed based on Mean Absolute Percentage Error (MAPE) [38] that can be defined as:

$$MAPE = \frac{1}{N} \sum_{t=1}^{N} \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

where $y_t$ and $\hat{y}_t$ are the actual and combined forecast value at some time period $t$ and $N$ is the number of observations used for performance evaluation and comparison.

C. Results and Discussion

Figure 3 displays $MAPE_{NN}$ values of $n_{total}$ NN models and $MAPE_{BMA}$ for the month of March 2011. It is obvious from this figure that good results are obtained by combining the forecasts using BMA approach. The $MAPE_{BMA}$ is less than 95% of the MAPEs of individual NN models.

The results of the MAPEs obtained using $BESTcomb$, $TRIMcomb$, $BMAcomb$ and the $Naive$ approach for the

\begin{table}[h]
\centering
\caption{Summary of parameters used in proposed model.}
\begin{tabular}{|l|l|}
\hline
\textbf{Model parameters} & \textbf{Settings} \\
\hline
Output layer transfer function & pure-linear \\
Hidden layer transfer function & tangent-sigmoid \\
Performance function & MSE with regularization \\
Epochs & 1000 \\
Performance measure & MAPE \\
$n_{total}$, $k_1$ & 100 \\
$n_{test}$ & 20 \\
AlphaOPT & the optimal value of $\alpha$ \\
$b$ & estimated coefficients with $BMAcomb$ \\
TUAhead & 48 \\
\hline
\end{tabular}
\end{table}
The performance of the aggregation algorithms of NN ensembles is evaluated with Australian electricity demand data. According to the result, the average MAPE of the BMAcomb leads to the best forecast. The BMAcomb enhances the performance by around 15% compared to Naive approach. The Bestcomb is able to improve its average error by 2% sharing almost similar results as Naive approach. However the TRIMcomb could not perform well in this research. That is the optimal amount of trimming in the TRIMcomb method did not produce good forecasting result. The reason may be, that the optimal trimming amount determined in this research work has trimmed greater amount of forecasts. This contrast the suggestion made by Goodwin [39] that greater amounts of trimming yield greater accuracy.

This research has focused on combination methods for an ensemble of neural networks. Advanced combination...
methods such as Frequentist model average [40] and other time varying combination methods may be investigated for combining NN ensemble. Aggregation of forecasts produced from hybrid models should be considered so as to get diverse forecasts. New designs of constructing NN ensemble or modification of existing designs should be consider to continue research in the field of ensemble modelling. Different alternatives of improving the performance of neural network model listed in [41], can be utilized in ensemble construction.

REFERENCES


